НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Інститут прикладного системного аналізу

Кафедра математичних методів системного аналізу

До захисту допущено:

В.о. завідувача кафедри

_____ Оксана ТИМОЩУК

«___»_____20__ р.

Дипломна робота

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Системний аналіз і управління»

спеціальності 124 «Системний аналіз»

на тему: «Модель зв'язку між швидкістю розповсюдження COVID-19 та регіональними міграційними потоками на прикладі США»

Виконав:
студент IV курсу, групи КА-63
Ніколаєнко Богдан Віталійович _____

Керівник:
асистент,
Макуха Михайло Павлович _____

Консультант з економічного розділу:
доцент, к.е.н., доцент кафедри ТТПЕ
Шевчук Олена Анатоліївна _____

Консультант з нормоконтролю:
доцент, к.т.н., доцент кафедри ММСА
Коваленко Анатолій Єпіфанович _____

Рецензент:
доцент, к.т.н
Харченко Константин Васильович _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ – 2020 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 124 «Системний аналіз»

Освітньо-професійна програма «Системний аналіз і управління»

ЗАТВЕРДЖУЮ
В.о.завідувача кафедри
_____ Оксана ТИМОЩУК
«___»_____20__ р.

ЗАВДАННЯ
на дипломну роботу студенту
Ніколаєнку Богдану Віталійовичу

1. Тема роботи «Модель зв'язку між швидкістю розповсюдження COVID-19 та регіональними міграційними потоками на прикладі США», керівник роботи Макуха Михайло Павлович, асистент, затверджені наказом по університету від «_25_»_травня_ 20_20_ р. №_1143-с__

2. Термін подання студентом роботи 08 травня 2020 року _____

3. Вихідні дані до роботи: відкриті набори даних про міграційні потоки населення США та дані про захворювання на COVID-19

4. Зміст роботи: огляд предметної області, розробка програмного продукту на основі обраної теми, аналіз результатів

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо)

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Економічний	к.е.н., доцент Шевчук О.А.		
Нормоконтроль	доцент, к.т.н. Коваленко А.Є.		

7. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Отримання завдання	12.04.2020	Виконано
2	Збір та обробка інформації	20.04.2020 – 26.04.2020	Виконано
3	Підготовка теоретичної частини диплому	27.04.2020 – 01.05.2020	Виконано
4	Обрання методів та алгоритмів для виконання задачі	01.05.2020 – 03.05.2020	Виконано
5	Розробка програмного продукту	05.05.2020 – 12.05.2020	Виконано
6	Виконання функціонально-вартісного аналізу	25.05.2020 – 27.05.2020	Виконано
7	Підготовка презентації	25.05.2020 – 30.05.2020	Виконано
8	Попередній захист дипломної роботи	01.06.2020 – 03.06.2020	Виконано
9	Захист дипломної роботи	15.06.2020 – 18.06.2020	Виконано

Студент Богдан НІКОЛАЄНКО

Керівник Михайло МАКУХА

РЕФЕРАТ

Дипломна робота: 73 с., 6 табл., 26 рис., 2 дод. та 21 джерело.

МОДЕЛЬ ЗВ'ЯЗКУ МІЖ ШВИДКІСТЮ РОЗПОВСЮДЖЕННЯ COVID-19 ТА РЕГІОНАЛЬНИМИ МІГРАЦІЙНИМИ ПОТОКАМИ НА ПРИКЛАДІ США.

Об’єкт дослідження – регіональні дані про пересування людей між штатами та дані епідеміологічні дані про розвиток епідемії COVID-19.

Предмет дослідження – сучасні моделі для аналізу та симуляції епідеміологічних процесів різного типу.

Мета роботи – розробка програмного продукту для виявлення зв’язку між швидкістю розповсюдження захворювання COVID-19 та внутрішніми міграційними потоками Сполучених Штатів Америки.

Актуальність – дослідження епідеміологічної сфери для практичного аналізу є важливою темою для всього людства, особливо за теперішньої ситуації в світі, коли більшість країн охопила пандемія коронавірусу.

У роботі був створений програмний продукт для аналізу розвитку епідеміологічного процесу на території США та за отриманих результатів був зроблений висновок про їх зв’язок з міграційними потоками.

Шляхи подальшого розвитку предмета дослідження – розширення факторів, які впливають на моделювання епідемій, збільшення точності та поглиблення рівня досліджень.

ABSTRACT

Diploma work: 73 p., 6 tabl., 26 fig., 2 appendix and 21 sources.

THE MODEL OF RELATIONSHIP BETWEEN COVID-19 PROPAGATION AND REGIONAL MIGRATION FLOWS BASED ON US CENSUS DATA.

The object of study – regional data on the movement of people between states and epidemiological data on the development of the COVID-19 epidemic.

The subject of research is modern models for analysis and simulation of epidemiological processes of different types.

Purpose – develop a software product to identify the relationship between the rate of spread of COVID-19 and internal migration flows of the United States.

Actuality – research the epidemiological field for practical analysis is an important topic for all humanity, especially in the current situation in the world, when most countries have been affected by the coronavirus pandemic.

A software product was created in the work for the analysis of the development of the epidemiological process in the United States and the results concluded that they are associated with migratory flows.

The further development of the research subject – the expansion of factors influencing the modeling of epidemics, increasing the accuracy and deepening of research.

ЗМІСТ

ВСТУП

Нещодавній спалах вірусу 2019-nCov в Ухані, Китай спричинив обвал фінансових ринків та світової економіки, та викликав неабияку паніку серед населення в усьому світі. 30 січня 2020 року на засіданні Всесвітньої організації охорони здоров'я, у Женеві пандемію вірусу 2019-nCov було визнано надзвичайною ситуацією в галузі охорони здоров'я міжнародного значення.

На момент написання цього документа ще не було виявлено ефективного лікування, яке б відповідало всім нормам та стандартам Всесвітньої організації охорони здоров'я. Поряд з цим досі залишаються невідомими важливі епідеміологічні фактори, такі як основна кількість репродукції (середня кількість людей, які можуть заразитись від середньостатистичного хворого на вірус 2019-nCov).

В наші часи, коли рівень глобального зв'язку та мобільності є безпрецедентним такі пандемії являють собою загрозу світового масштабу для населення всієї планети. Можна припустити, що за нинішнього темпу розвитку суспільства подією глобальної катастрофи, коли кількість жертв по всій планеті перевищує 100 мільйонів, може стати саме пандемія такого типу, а не ядерна чи кліматична катастрофа. З кожним роком темпи урбанізації по всьому світі зростають, а наші густонаселені, динамічні міста перетворюються в потенційні вузли розповсюдження небезпечних захворювань.

Мета цієї роботи – розробити модель, яка допоможе знайти зв'язок між поширення вірусу 2019-nCov та пересуванням населення на прикладі Сполучених Штатів Америки, що дозволить зрозуміти, на які потоки слід звертати увагу, для запобігання розповсюдження вірусу.

РОЗДІЛ 1 ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Загальні поняття про сімейство Коронавірусів

Перш за все визначимося з поняттям Коронавірус.

Термін «Коронавірус» об’єднує в собі велику групу вірусів, здатних вражати птахів, ссавців та людей. Група отримала назву «Коронавіруси» внаслідок незвичайної будови молекули вірусу, яка має виступи, що нагадують точки на короні.

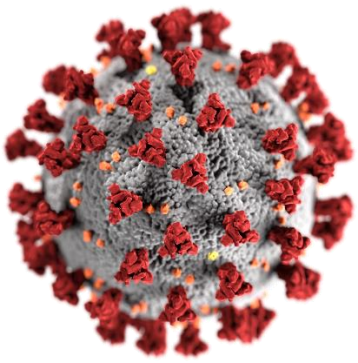


Рисунок 1.1 – Модель молекули коронавірусу [1]

Це сімейство вірусів відносять, до категорії найнебезпечніших для людини, через їхню будову та вплив на організм. Генетичний матеріал типових представників даного виду представлений молекулами нуклеїнової кислоти. Під час зараження вірусом молекули РНК (рибонуклеїнова кислота) прикріплюються до клітин інфікованого та створюють власні копії, таким чином поширюючись в організмі. Якщо копію не вдається створити, то відбувається мутація – зміна молекул РНК. Здатність вірусів даного типу до постійних мутацій, перешкоджає швидкій розробці ефективної вакцини, для запобігання епідемій, оскільки ці зміни відбуваються випадковим чином.

Дана група має близько 100 видів вірусів. Ці види відрізняються по фактору ризику та тяжкості захворювання. Наприклад, такий штам, як MERS-CoV, має рівень смертності більше 30% від кількості всіх інфікованих. В залежності від виду вірусу, хвороба може протікати по різному – це може бути, як звичайна застуда з її звичними симптомами(біль в горлі, лихоманка), так і пневмонія, бронхіт. В 2003 році був відкритий новий вид коронавірусу людини, який мав особливий патогенез. Даний штам вірусу перетікав в дуже тяжку форму, та викликав тяжкий гострий респіраторний синдром, через що і отримав назву SARS-CoV.

1.2 Спалахи епідемій коронавірусу

Серед великої кількості штамів коронавірусу, лише 7 мають вплив на людину, серед яких 3 протікають в тяжкій формі.

1.2.1 SARS-CoV

SARS-CoV або ж тяжкий гострий респіраторний синдром, РНК-вірус сімейства коронавірусів, який був виявлений в 2003 році в провінції Гуандун, Китай. Даний штам вірусу може протікати в тяжкій формі, з незначними симптомами на початковій стадії. В тяжких формах спостерігається атипова пневмонія та грип. В 2003 році, під час спалаху атипової пневмонії, від вірусу SARS-CoV померло близько 9% від всіх інфікованих, в людей, старших 60 років, смертність досягала 50%.

Серед симптомів виділяли головну біль, лихоманку, діарею, слабкість. В той же час, досі ніхто не виділив специфіку проявлення симптомів та базову групу для чіткого визначення захворювання, адже в пацієнтів, в залежності від віку чи імунної системи, вони проявлялися з різною тяжкістю та в різних періодах протікання хвороби, що сприяло швидкому розповсюдженню вірусу.

Передача вірусу відбувалась від людини до людини, також як носіїв виділяли приматів та кажанів, серед яких також спостерігалось активне протікання хвороби. Під час спалаху в 2003 році SARS-CoV було розповсюджено на 26 країн, в яких було зареєстровано більше 8000 інфікованих. Після завершення епідемії захворювання на цей штам фіксуються періодично, хоч і з незначною кількістю інфікованих. До сьогодні ефективного методу профілактики від SARS-CoV не було виявлено.

1.2.2 MERS-CoV

MERS-CoV – коронавірус, який викликає гострий респіраторний синдром, має широке поширення на Близькому Сході, зокрема в країнах Аравійського пів острову. Даний штам вірусу є зоонозним, тобто притаманним середовищем існування для нього – є тварини, але за особливих ситуацій чи мутацій, він здатен передаватись від тварин до людини. Теорія передачі даного штаму від людини до людини досі немає вагомих аргументів для існування. Осередком існування цього типу коронавірусу є верблюди, шо пояснює його широке розповсюдження в країнах Близького Сходу, особливо в Саудівській Аравії та Йорданії.

MERS-CoV вперше було зареєстровано в 2012 році в Саудівській Аравії, тоді він створив спалах епідемії серед верблюдів, яка поступово поширювалась на людей та розповсюджувалась. З неабиякими труднощами,

пов’язаними з бюрократичними незгодами, вченим вдалось виділити матеріали вірусу, щоб ізолювати його від людей та тварин, що знизило темпи розповсюдження, але постійні мутації коронавірусу, спричиняли нові спалахи захворювань в 2013, 2015 роках, щоправда темпи розповсюдження та смертності були нижчими, в порівнянні з епідемією 2012 року, смертність якої складала більше 30%. Ефективних ліків від цього вірусу немає, але вченими активно ведеться розробка вакцини, яка буде здатна корегувати рівень інфікування, тобто створить можливість блокувати канали передачі вірусу між тваринами та тваринами і людиною.

1.2.3 SARS-CoV-2

SARS-CoV-2 – РНК-вірус, відомий також як 2019-nCov, викликає тяжкий гострий респіраторний синдром коронавірусу, який в свою чергу перетікає в хворобу, яка отримала назву COVID-19. Даний штам коронавірусу був виявлений в Ухані, Китай в грудні 2019 року. Під час дослідження тяжкої форми атипової пневмонії в рибонуклеїновій кислоті пацієнта був виявлений невідомий патоген. Одразу після цього почали з’являтись нові повідомлення про тяжкі форми захворювань на вірусну пневмонію, причиною яких є невідомий збудник. Вірус передається як між людьми, так і від тварини до людини, що тільки збільшує темпи поширення інфекції. На початкових етапах розповсюдження коронавірусу, основним місцем інфікування, стали ринки морепродуктів в Китаї, де відбувався продаж живих тварин, за теоріями деяких вчених це і стало поштовхом до зростання темпів інфікування. Поступово вірус розповсюджувався по всіх провінціях Китаю, а кількість інфікованих почала ставати вдвічі більшою кожних 7 днів. За такої динаміки росту Всесвітня організація охорони здоров’я прийняла рішення, про введення

надзвичайного стану міжнародного значення, що потім змінилося на статус пандемії.

Завдяки своїй будові SARS-CoV-2 починає еволюціонувати та створювати нові підтипи вірусу. Зараз виділять два модифіковані види – L, та S. Підтип L є більш агресивним збудником, протікання хвороби такого вірусу відбувається, в більшості випадків, в тяжкому стані. Окрім цього коефіцієнт репродукції цього підтипу, тобто кількість людей, що можуть заразитись від особи, яка є інфікованою, є більшим, що сприяло швидкому рознесенню вірусу на ранніх етапах. Другий підтип – S, зустрічається набагато рідше, та є менш агресивним, але, незважаючи на це, він поступово витісняє SARS-CoV-2 типу L, що сприяє зменшенню смертності. Така поведінка пояснюється більш сильнішим селективним тиском.

Як і притаманно сімейству коронавірусів хвороба, викликана вірусом SARS-CoV-2 може протікати в різних ступенях тяжкості. Інкубаційний період триває від 1 до 14 днів, що дозволило широко рознести вірус по всьому світі, вже за місяць після виявлення хвороби, кількість інфікованих за межами Китаю значно перевищувало кількість інфікованих по країні. Перші симптоми від захворювання можуть нагадувати звичайну простуду або лихоманку та в подальшому протікати без переходу в тяжку форму. В більшості випадків хвороба протікає в легкій формі, але в деяких випадках можуть проявлятись симптоми атипової пневмонії, пневмонії з гострим респіраторним дистрес-синдромом, або ж, навіть запальні процеси по типу цитокінового шторму, який викликає який викликає хаотичну реакцію імунної системи на збудник хвороби, та робить організм нездатним до боротьби з вірусом, наслідки такого процесу, в більшості випадків, є летальними.

Не зважаючи на, зусилля всієї світової спільноти ефективного способу боротьби з хворобою COVID-19 поки не виявлено. Всі винайдені методи дають або частковий та недовготривалий ефект, або ж діють лише на дуже вузьку вибірку хворих, не блокуючи основного збудника хвороби.

1.3 Прогнозування вірусних захворювань

За всю свою вікову історію населення планети пережило велику кількість епідемій, котрі в ті чи інші часи закінчувались масовою смертністю. Перед людством постало питання, як контролювати епідеміологічні процеси, чи, навіть, як їх контролювати. Починаючи з XX століття, коли збір та фіксація даних перейшли на новий рівень, та з розвитком математики та статистики стало можливим робити деякі припущення про період та піки протікання захворювань. В останні роки число робіт по прогнозуванню епідемій стрімко зросло, це зумовлено чітким збором даних всіх необхідних факторів, що стосуються появи та існування вірусів.

Епідеміологічні моделі бувають різних типів та призначень. Одні можуть показати короткостроковий прогноз, який буває необхідний для виявлення так званих ревізій, тобто раптових поновлень спалахів вірусу. В той же час довгострокові прогнозі використовують для огляду картини розповсюдження вірусу в цілому, для знаходження піків захворювань або оцінки кількості необхідних лікарських препаратів для запобігання епідемій. Маючи такий великий спектр використання довгострокові прогнози мають і вагомий недолік, у вигляді неточності прогнозу, яку ніяким чином неможливо значно збільшити. Незважаючи на цей достатньо вагомий фактор оцінки моделей довгострокові прогнози активно використовуються в епідеміологічних моделях через своє важливе стратегічне значення. Важливим фактором при прогнозуванні захворювань чи розвитку епідемій є вибір моделі та підходу до реалізації. Як відомо на різних етапах хвороба може поводити себе по різному, тому беручи це до уваги слід вірно обирати техніку прогнозування.

Прогнозування починається з даних. Дані в області вірусології є різнотипними, в залежності від типу хвороби, її тяжкості чи масштабності,

можуть бути важливими для побудови моделі ті чи інші фактори. Об’єднує всі дані про захворювання час – це один з основних факторів, на який звертають увагу при всіх дослідженнях. Саме від часу, чи періоду певної типової поведінки вірусу роблять висновки про його природу, степінь небезпеки. Саме від того з якою частотою відбувається збір даних залежить, наскільки точною вдасться розробити модель прогнозу.

1.4 Методи прогнозування вірусних захворювань

Існує велика кількість моделей та методів для прогнозування вірусних захворювань, але всі вони мають свою специфіку використання, свої переваги та недоліки, тому в даному розділі ми розглянемо основні, найбільш вживані та ефективні методи.

1.4.1 Регресійні моделі

Регресійні моделі – є одним з найпопулярніших методів прогнозування вірусних захворювань. Основною задачею регресії являється знаходження функціональної залежності показниками захворюваності та факторами, які впливають на захворюваність, задля чого відновлюються та формуються невідомі параметри. Основними видами в дослідженні вірусних інфекцій є адаптивні та неадаптивні регресійні моделі. Вони можуть застосовуватись під час дослідження різних типів захворювань, задля визначення інформації, яка буде доступна тільки конкретному виду моделі.

Адаптивні регресійні моделі використовують здебільшого вже під час протікання хвороби, задля прогнозування та аналізу її подальшої поведінки. Адаптивні моделі використовують так зване поняття ширини вікна, що відображає число випадків проявлення збудника, які використовуються для оцінки прогнозу. В залежності від ширини вікна змінюється точність та складність моделі: при збільшенні ширини вікна, точність знижується, це пояснюється тим, що зростає складність функції, необхідної для обробки більшого потоку даних. Ефективність даного типу в виявлені всіх локальних коливань епідеміологічних показників.

Особливістю неадаптивних моделей є можливість врахувати всі попередні випадки захворювань. Такі моделі використовують всі накопичені дані для побудови прогнозу, а також дані показників схожих вірусів. Цей тип прогнозу здебільшого ефективний в використанні для передбачення сезонних видів вірусів. Яскравим прикладом такої моделі є модель Серфлінга для виявлення та прогнозування нових спалахів епідемій з чітко виявленими сезонними факторами:

$$\widehat{y_t} = \gamma \mathrm{e}^{\sum_{j=0}^{v} a_j t^j + \sum_{j=0}^{k} (\beta_{2j-1} sin\theta_j + \beta_{2j} cos\theta_j) + \rho_1 x_t^1 + \rho_2 x_t^2 + \cdots}$$

Змінні x_t^i – приймають значення тих факторів, що впливають на швидкість та степінь тяжкості захворювання, змінні γ, ρ_j являються параметрами, котрі змінюються для кожного типу вірусу [1-3].

1.4.2 Баєсівські мережі

Даний тип моделювання епідеміологічних процесів ґрунтується на відображенні моделі у вигляді орієнтованого графа, на вершинах якого розташовуються змінні моделі, а ребрам відповідають ймовірнісні залежності між цими показниками. В результаті машинного навчання модель такого типу здатна оцінювати з якою ймовірністю трапиться та чи інша подія, при певній послідовності явищ. В вірусології Баєсівські мережі активно стають популярними саме у вигляді простої форми прихованих марковських моделей. Зручність прихованих марковських моделей в тому, що їх можливо реалізувати як великій, так і при малій кількості даних, про протікання вірусу, але через це одразу з'являється недолік у вигляді складності моделі для довгострокових прогнозів, тому приховані марковські мережі ефективно використовують тільки для короткотривалих прогнозів. Найчастіше такі моделі використовуються для виявлення різкого збільшення темпів розповсюдження інфекцій [4].

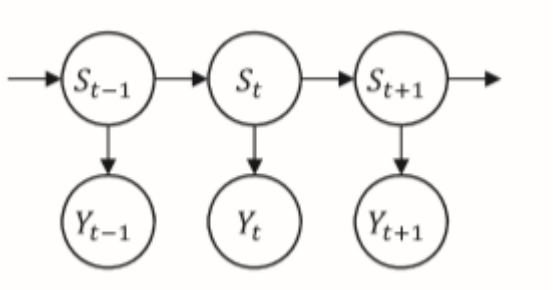


Рисунок 1.2 – Схематичне зображення залежності в прихованих марківських мережах [2]

Ідея моделі прихованих марковських мереж базується на співставленні кожній випадковій величині Y_t випадкову величину S_t, яка в даний момент не спостерігається, але визначає умовний розподіл Y_t. В даному випадку, в якості

параметру Y_t може виступати число звернень до медичних закладів, чи повідомлень про захворювання, в такому випадку S_t буде відповідати загальній кількості інфікованих. Таким чином можливо проаналізувати параметри моделі в конкретний момент часу t, задавши розподіл для цих двох змінних, враховуючи те, що Y_t може залежати тільки від S_t, яка, в свою чергу, може залежати тільки від свого попереднього значення [5].

1.4.3 Експоненційне згладжування

В деяких моделях дані з часовими мітками аналізують, сприймаючи їх за випадковий процес, який видає певний сигнал та шуми. В таких задачах ставиться за пріоритет позбавлення від шумів, задля виділення якомога чіткішого сигналу, котрий і буде відображати епідеміологічний стан протікання захворювання. Саме такий підхід використовується при експоненційному згладжуванні.

В експоненційному згладжуванні рівень захворюваності представляється у вигляді зваженої суми певної кількості останніх спостережені:

$$l_t = \alpha y_t + (1 - \alpha) l_{t-1}$$

За такою формулою прогноз захворюваності y_t, в момент часу t визначається по принципу $\hat{y}_{t+h} = l_t$, де l_t є згладженим значенням, а коефіцієнт α відповідає за нормування та зменшення ваги застарілих даних. Цей підхід раціонально використовувати, коли епідемії не мають чітко виявленої сезонності або тренду. Для сезонних захворювань використовують так зване потрійне експоненційне згладжування.

Яскравим прикладом потрійного експоненційного зважування є адитивна сезонна модель Хольтера-Вінтерса:

$$l_t = \alpha(y_t - s_{t-T}) + (1-\alpha)(l_{t-1} + r_{t-1}),$$
$$r_t = \gamma(l_t - l_{t-1}) + (1-\gamma)r_{t-1},$$
$$s_t = \delta(y_t - l_t) + (1-\alpha)s_{t-T},$$
$$\hat{y}_{t+h} = l_t + hr_t + s_{t-T+h},$$

де r_t відповідає за тренд, s_t – певний фактор сезонності , T – тренд сезону, γ, δ – коефіцієнти згладжування.

Перевагою даної моделі є властивість не реагувати на будь-які випадкові відхилення, пов’язані з певним періодом чи сезоном, які залежать від показників захворюваності.

Методи прогнозування, основані на експоненційному згладжуванні здебільшого є популярними в сфері економіці, та не так часто використовуються для прогнозування захворювань, хоча можуть давати досить непогані результати в моделях як для довгострокового періоду, так і короткострокового [6].

1.4.4 Калманівські фільтри

Фільтри Калмана користуються широкою використовуються в області інженерії та економетрики, і тільки починають набувати популярності в прогнозуванні епідеміологічних процесів. Даний тип моделей має певну схожість в представленні стані з, раніше розглянутою моделлю експоненційного згладжування та моделлю авторегресії. В загальному вигляді епідеміологічні процеси для нашої моделі можна записати в такому вигляді:

$$x_t = Ax_{t-1} + w_t,$$
$$y_t = Hx_t + Df_t + v_t,$$

де x_t – вектор змінних стану системи, в певний момент часу t, вектор спостережень – y_t, f_t – відповідає значенням зовнішніх факторів, що впливають на систему, w_t, v_t – шуми. В залежності від типу того, прогноз довгостроковий чи короткостроковий матриці A, H, D, які визначають епідеміологічний процес, можуть змінюватись [7].

Завдяки такій формі представлення стає можливим створення узагальнюючої моделі поширення вірусів, яку, на основі методів теорії регулювання, можна використовувати для прогнозування протікання захворювань.

1.5 Висновок до розділу 1

В даному розділі ми розглянули основні поняття, які більш ширше розкривають для нас вірусологію, а саме частину таких небезпечних видів інфекцій, як коронавіруси. Тільки повертаючись назад та розглянувши історію спалахів цієї хвороби, з'ясувавши певні аспекти будови її молекул та процесу зараження ми зможе обрати вірну стратегію для побудови моделі прогнозу. Саме для цього ми також розглянули найбільш відомі моделі для створення прогнозів протікання захворювань, щоб з застосуванням їхніх ідей та структури, в певному комбінуванні з новими, маловідомим методиками спробувати досягти суттєвого зменшення похибок у власних моделях прогнозування.

Створення достатню точних моделей для прогнозування епідемій є однією з цільових задач для всього людства в теперішні дні. За нинішньої

ситуацій, коли саме природа стає суттєвою загрозою для здоров’я людей, потрібно створити моделі високої точності для прогнозування або імітації епідемій вірусів з якими людство вже стикалось та, навіть з тими, які були нещодавно відкриті. Як нам відомо з плином часу починають відкриватись нові збудники захворювань, реакціях яких на людину досі невідома, від чого може виникнути серйозні проблеми в майбутньому.

Поряд з цими рівень урбанізації та темпи соціальної активності з року в рік постійно збільшується, це означає що щільність населення та контактування людей між собою постійно збільшується. Як показує практика, а саме нинішня ситуація в світі, ці фактори є не малозначними при розгляді моделей прогнозування епідемій. Тож в цій роботі ми будемо намагатись віднайти як саме буде впливати збільшення чи зменшення соціальних контактів на швидкість розповсюдження вірусів, на прикладі міграційних потоків.

РОЗДІЛ 2 ОГЛЯД МАТЕМАТИЧНИХ МОДЕЛЕЙ ТА АЛГОРИТМІВ

2.1 Вступ до розділу 2

В даному розділі ми детально розглянемо моделі, які будемо використовувати в нашій роботі та їх математичну формалізацію. Для імітації та подальшого аналізу протікання епідемій була обрана модель SIR.

Такі моделі, як правило, досліджені за допомогою звичайних диференціальних рівнянь (які є детермінованими), але також можна розглядати в стохастичною структурою, яка є більш реалістичним, але і більш складним для аналізу.

Полігамні моделі можуть бути використані для прогнозування властивостей, як хвороба поширюється, наприклад, поширеність (загальне число інфікованих) або тривалість епідемії. Крім того, модель дозволяє зрозуміти, яким чином різні ситуації можуть вплинути на результат епідемії, наприклад, що найбільш ефективним методом є для видачі обмеженої кількості вакцин в даній популяції.

2.2 Модель SIR

На сьогоднішній день SIR-моделі є одними з найпопулярніших в області моделювання інфекційних захворювань, завдяки великій варіативності форм моделі та легкості в пристосуванні під різні форми та етапи епідеміологічних процесів. Дана модель за допомогою використання диференційних рівнянь описує зміни певних трьох показників вірусу, з чого і пішла назва моделі:

– S – кількість сприйнятливих до вірусу осіб, які не заражені, але могли заразитися.
– I – кількість заражених. Ці особи мають захворювання і можуть передавати його вразливим до вірусу особам.
– R – кількість вилучених особин. Вони можуть мати або не мати захворювання, але вони не можуть заразитися, і вони не можуть передати хворобу іншим. У них може бути природний імунітет, або вони, можливо, одужали від хвороби і не застраховані від її повторного зараження, або у них може бути захворювання, але вони не здатні передати його (наприклад, тому, що вони моглибути поміщені в ізоляцію), або вони можуть померли.

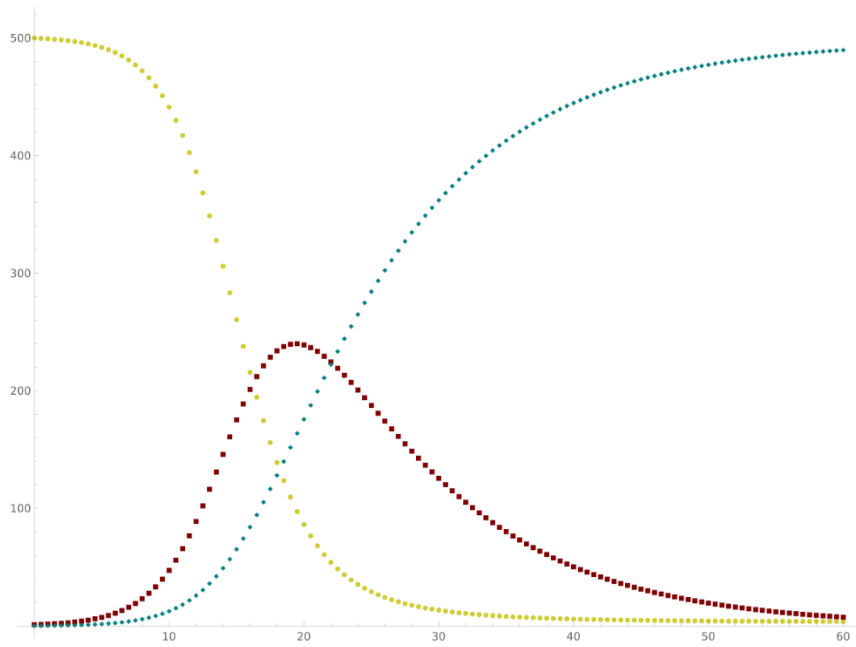


Рисунок 2.1 – Загальний вигляд моделі SIR [3]
жовта пряма – S, червона – I, блакитна – R

Під час контактування з інфікованими особами, ті, хто здатен реагувати на вірус, можуть перейти до групи інфікованих, а інфіковані, в свою чергу, можуть потрапити в групи осіб, які позбулись вірусу, внаслідок здобуття імунітету через одужання, або ж смерті причиненої захворюванням.

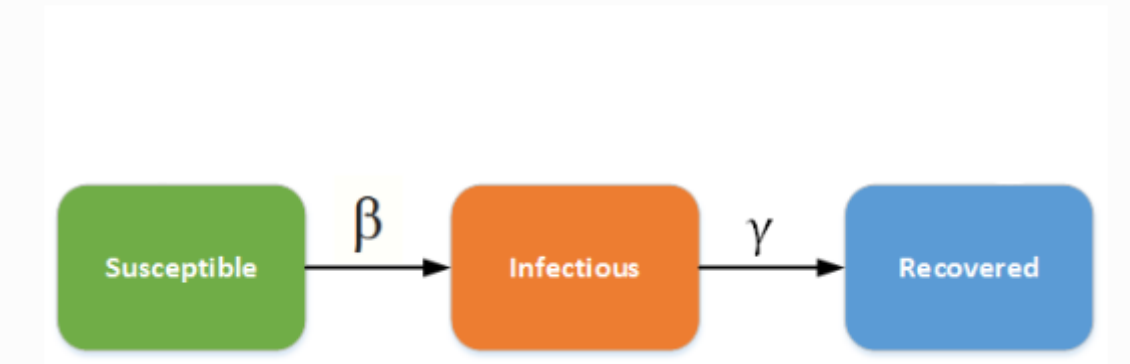


Рисунок 2.2 – Цикл зміни групи в моделі SIR [4]

В загальному вигляді SIR-модель представляють у вигляді системи трьох диференційних рівнянь. Першого рівняння системи описує швидкість зміни чисельності осіб, які можуть реагувати на вірус, друге – кількість інфікованих, третє – осіб, які вже не є переносниками вірусу [12].

2.2.1 Модель SIR без врахування життєвої динаміки

Деякі вірусу відрізняються значно вищою динамікою епідемії порівняно з показниками смертності та народжуваності, тому зазвичай ці фактори не враховуються при побудові моделі, за що вона і отримала свою назву. Як і будь-який підвид моделі SIR, дана модель SIR, без врахування життєвого циклу описується системою диференційних рівнянь. Представлена нижче система була створена на основі теорії Кермак-Маккендрика та являє собою систему нелінійних рівнянь, а отже аналітичний розв’язок можливо отримати лише в неявному вигляді [8-9].

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta IS}{N}, \\ \frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \\ \frac{dR}{dt} = \gamma I, \end{cases}$$

Коефіцієнт $\beta > 0$ визначає ймовірність того, що особа, яка здатна реагувати на вірус при контакті з інфікованим може заразитись. Коефіцієнт $\gamma > 0$ визначає, з якою швидкістю вдається вилікуватись, а $1/\gamma$ – середня протяжність протікання хвороби. Окрім цього слід звернути увагу на те, що:

$$S(t) + I(t) + R(t) = const = N$$

В даному випадку константа N відображає загальну кількість населення. Інакше кажучи, це означає, що ми розділяємо все населення на три групи, які змінюють свій розмір з плином часу, але загальна кількість завжди залишається незмінною. Кількість нових інфікованих залежить зокрема від таких показників : β, S, I. Це означає, що ми отримуємо нових хворих та показники швидкості захворювання за одиницю часу під час однорідного змішування інфікованих та чутливих класів. Оскільки:

$$\mathrm{R(t)} = \mathrm{N} - \mathrm{S(t)} - \mathrm{I\,(t)},$$

то систему можна звести до системи з двох однорідних рівнянь. Нехай кожна інфікована особа має k контактів, кожен з яких є достатнім для передачі інфекції, та k є незалежним від популяцій. Тоді k * S / N цих контактів відбувається із сприйнятливими людьми. Таким чином, якщо певна фракція τ, чутливих до хвороби осіб, знаходиться в групі інфікованих та має достатній для передачі контакт з ними, то будь-яка інфікована особа інфікує k * τ * S / N

сприйнятливих осіб за одиницю часу. Таким чином β = b / N, де b = k * τ. Параметром τ називають показником передачі інфекційного захворювання.

2.2.2 Число репродукцій R

Репродуктивне число є одним з найважливіших показників в розробці епідеміологічних моделей. В загальному розумінні термін «репродуктивне число» або число репродукцій означає середню кількість нових заражень, викликаних однією інфікованою особою, яка мала достатнім для зараження контакт, з чутливими до хвороби класами. В реальному житті це значення залежить від багатьох факторів, таких як географічна область протікання хвороби, клімат чи густота населення. Для певних захворювань, які відомі або з якими людство стикається періодично, таких як грип, поліомієліт чи хвороба, яку спричинює вірус Ебола, число репродукцій відомо. Ми розглянемо математичне представлення для числа репродукцій хвороби, викликаної вірусом 2019-nCoV.

Визначимо число репродукцій:

$$R_0 = \frac{\beta}{\gamma}$$

Або ж це відношення середньої швидкості зараження інфікованим, чутливу до хвороби особу, до степені одужання, або ж середнього часу протяжності хвороби.

З іншого боку дану формулу можна записати, як:

$$R_0 = \frac{T_r}{T_c}$$

В якості T_c тут виступає середня швидкість між контактами - β^{-1}, а в якості T_r середня швидкість до одужання γ^{-1}.

Репродуктивне число використовується для визначення кількості осіб, які входять до чутливого класу, це яскраво видно, якщо провести деякі математичні перетворення з системою диференційних рівнянь. Створивши відношення між першим та третім рівнянням та провівши інтегрування, ми отримуємо рівняння виду:

$$S(t) = S(0)e^{-\frac{R_0(R(t)-R(0))}{N}},$$

де $S(0)$та $R(0)$ виступають в якості значень класів чутливих до вірусу та осіб, які позбулись вірусу, на початок епідемії [10].

Як ми можемо бачити кількість чутливих класів в конкретний момент часу напряму залежить від числа репродукцій та може змінюватись з експоненційною швидкістю. Тобто якщо число репродукцій $R_0 = 3$, це означає, що кожна інфікована особа буде передавати хворобу трьом чутливим індивідам, кожен з яких, в свою чергу, заражає трьох нових. Схематичне зображення такого варіанту представлене на Рисунку 2.3.

В класі інфікованих осіб репродуктивне число займає також важливу роль, адже, як видно з формули:

$$\frac{dI}{dt} = (R_0 \frac{S}{N} - 1)\gamma I$$

Епідеміє відбудеться лише за умови:

$$R_0 > \frac{N}{S(0)}$$

Адже саме тоді Початкове число інфікованих буде більше нуля, що і спричиняє розповсюдження хвороби [11]. За інших обставин, коли:

$$R_0 < \frac{N}{S(0)}$$

Маємо випадок, коли, в незалежності від початкового числа інфікованих, вірус не зможе розповсюдитись.

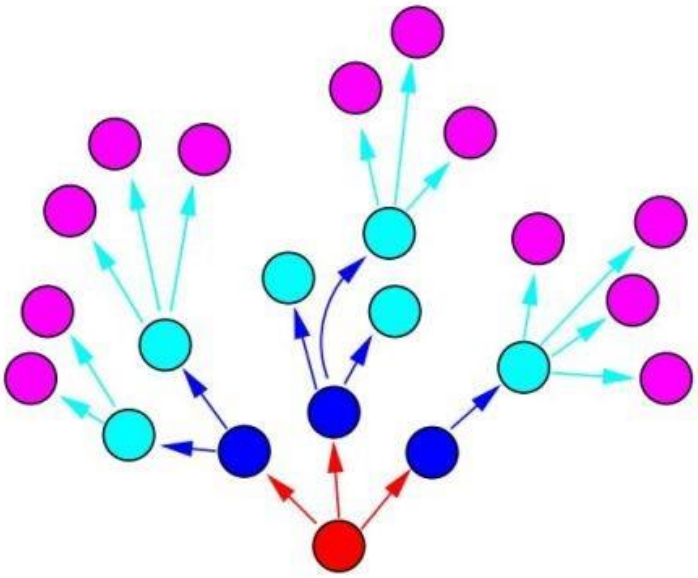


Рисунок 2.3 – Експоненційний ріст заражених особин на початку епідемії [5]

2.3 Модель SEIRS з врахуванням життєвого циклу

Модель SEIRS виступає в якості покращеної версії моделі SIR, та дозволяє скорегувати епідеміологічну модель під певні види захворювань. Існує велика кількість захворювань, які мають певний, так званий, «латентний» період протікання хвороби, під час якого вірус ніяк себе не проявляє, але поступово все глибше проникає в організм людини. Тому дана модель поєднує в собі одразу чотири етапи протікання хвороби, що забезпечує врахування окремого класу безсимптомного, але небезпечного для інших протікання хвороби.

Як можна було зрозуміти з назви, модель SEIRS з врахуванням життєвого циклу складається з чутливого до інфекцій класу – S, тих, хто вже був заражений, але не відчуває симптом та не може інфікувати інших – E, індивіди які є інфікованим та можуть передавати вірус чутливим до нього особам – I, та тих, хто вже не є переносником вірусу – R.

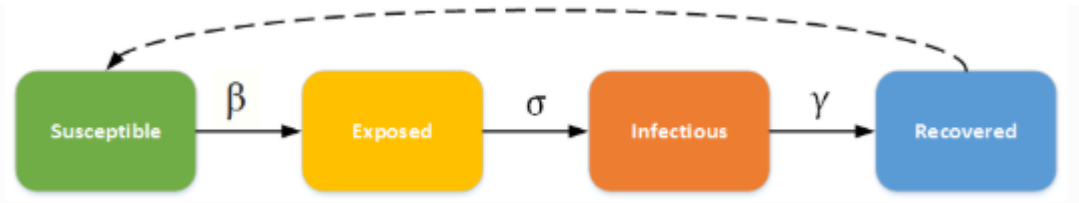


Рисунок 2.4 – Цикл зміни груп в моделі SEIR [6]

Модель SEIR враховує нерівність параметрів народжуваності та смертності окрім цього ми можемо враховувати ймовірність повторного зараження вірусом після одужання, а це є важливою умовою при дослідженні коронавірусу. На даний момент немає чіткої відповіді, про здобуття імунітету від хвороби COVID-19, але наша імунна система не здатна ефективно реагувати на повторне захворювання від чотирьох інших штамів цього вірусу,

тож певна ймовірність такого випадку є. В даній моделі, як можна зрозуміти зі схеми Рисунку 2.3, протікання хвороби кваліфікується за певним алгоритмом. Спочатку, при взаємодіє чутливого до захворювання класу, з інфікованим індивідом відбувається передача інфекції, після чого інфікована особа переходить до наступного класу. Після інкубаційного періоду особа починає проявляти симптоми захворювання та стає переносником вірусу. Після подолання вірусу, особа переходить до класу вилікуваних, але з певною ймовірністю може згодом знов потрапити до класу чутливих до вірусу осіб.

В загальному випадку модель SEIRS з врахуванням життєвого циклу записується у вигляді системи чотирьох диференційних рівнянь:

$$\frac{dS}{dt} = bN - \frac{\beta SI}{N} + \alpha R - dS,$$
$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E - dE,$$
$$\frac{dI}{dt} = \sigma E - \gamma I - dI,$$
$$\frac{dR}{dt} = bN + \gamma I - dR - \alpha R,$$

де α – виступає, як показник швидкості, з якою особи, які одужали, втрачають тимчасовий імунітет до вірусу; $\beta -$ відображає швидкість, з якою чутливі до хвороби особи можуть бути інфіковані та перейти до класу безсимптомного протікання хвороби; $\sigma -$ показує швидкість, з якою індивід переходить до класу інфікованих; $\gamma -$ швидкість, з якою відбувається протікання хвороби, тобто, як швидко протікає період від потрапляння до класу інфікованих, до переходу в клас одужавших [13].

Загальна кількість населення для даної моделі може змінюватись з часом:

$$S(t) + E(t) + I(t) + R(t) = N(t)$$

Число репродукцій для моделі SEIR відрізняється від числа репродукцій для моделі SIR, в даному випадку воно має вигляд:

$$R_0 = \frac{\alpha}{d + \alpha} \frac{\beta}{d + \gamma}$$

Як ми бачимо з формули чисельні зміни популяції, внаслідок коливання показників народжуваності та смертності, напряму впливають на репродукційне число.

Вагомим недоліком епідеміологічної моделі SEIRS є не можливість враховувати фатальні випадки протікання захворювання. В даному випадку ми розглядаємо лише можливість особи знову бути інфікованою, після потрапляння до класу Recovered, або ж залишитись в класі Susceptible. Але при цьому ми впускаємо велику частку випадків, коли індивід не може потрапити до будь-якого з цих класів, через те, що захворювання привело до фатальних наслідків. Саме для цього ми будемо використовувати модель SIR-F [14-15].

2.4 Модель SIR-F

Модель SIR-F є вдосконаленою версією раніше згаданих SIR та SEIRS моделей. Використання даної моделі дає можливість у врахуванні фактично підтверджених летальних випадків протікання захворювання. Цей фактор є важливим, адже при аналізі протікання хвороби ми можемо виокремити більш детально групу індивідів, які можуть бути ще раз інфіковані хворобою. В такий спосіб, маючи більший спектр задіяних груп інфікованих, ми можемо

отримати кращий аналіз показників вірусу, що, в свою чергу, посприяє покращенню прогнозування майбутньої епідеміологічної ситуації, за такими групами: кількість чутливих до хвороби; кількість інфікованих; кількість особин, які вилікувались; кількість летальних випадків.

Цикл переходу з однієї групи до іншої зображений на Рисунку 2.4

$$\begin{aligned} \mathrm{S} &\xrightarrow{\beta I} \mathrm{S}^* \xrightarrow{\alpha_1} \mathrm{F} \\ \mathrm{S}^* &\xrightarrow{1-\alpha_1} \mathrm{I} \xrightarrow{\gamma} \mathrm{R} \\ \mathrm{I} &\xrightarrow{\alpha_2} \mathrm{F} \end{aligned}$$

Рисунок 2.5 – Зміна групи в моделі SIR-F [7]

Відповідно до Рисунку 2.4, визначаємо, що S, I, R, як раніше і було зазначено – чутливі класи, інфіковані, ті хто вилікувався. Підгрупи визначаються, як S^* – кількість усіх зафіксованих випадків, а F – кількість фатальних випадків, які було підтверджено. При даному записі коефіцієнт α – виступає, як показник рівня потрапляння до групи F, від своїх класів; $\beta -$ відображає швидкість, з якою чутливі до хвороби особи можуть бути інфіковані та перейти до класу безсимптомного протікання хвороби; $\gamma -$ швидкість, з якою відбувається протікання хвороби.

Загальний вигляд модель SIR-F, у вигляді диференційних рівнянь:

$$\frac{dS}{dT} = -\frac{\beta SI}{N},$$

$$\frac{dI}{dT} = \frac{(1-\alpha_1)\beta SI}{N} - (\gamma + \alpha_2)I,$$

$$\frac{dR}{dT} = \gamma$$

$$\frac{dF}{dT} = \frac{\alpha_1 \beta SI}{N} + \alpha_2 I$$

де $N = S + I + R + F$ і є загальною кількістю населення, а T – час від початку епідемії.

З огляду на розглянуті види епідеміологічних та їхній розподіл за класами, можемо зробити висновок, що дана модель найбільш підходить для аналізу вірусу типу SARS-nCoV-19, аналізу поставленої задачі, та для розгляду майбутніх випадків та аналізу поведінки вірусу. Розглянемо більш детально модель [16].

Для корегування алгоритму під певні фактори розглянемо безрозмірний запис моделі, який будемо використовувати при аналізі.

Нехай:

$$(S, I, R, F) = N \times (x, y, z, w)$$
$$(T, \alpha_1, \alpha_2, \beta, \gamma) = (\tau t, \theta, \tau^{-1}\kappa, \tau^{-1}\rho, \tau^{-1}\sigma)$$

де N – загальна кількість населення, а τ – певна константа.

Тоді:

$$\frac{dx}{dt} = -\rho xy$$
$$\frac{dy}{dt} = \rho(1-\theta)xy - (\sigma + \kappa)y$$
$$\frac{dz}{dt} = \sigma y$$
$$\frac{dw}{dt} = \rho\theta xy + \kappa y$$
$$0 \leq (x, y, z, w, \theta, \kappa, \rho, \sigma) \leq 1$$

Відповідно число репродукцій для даної моделі буде:

$$R_0 = \rho(1-\theta)(\sigma+\kappa)^{-1} = \beta(1-\alpha_1)(\gamma+\alpha_2)^{-1}$$

Для заданої моделі розглянемо її графічне представлення (Рисунок 2.5), для значень: $R_0 = 2.5,\ \theta = 0.002,\ \kappa = 0.005,\ \rho = 0.2$, а початкові значення відповідних класів: $(x_0, y_0, z_0, w_0) = (0.999, 0.001, 0, 0)$ [21].

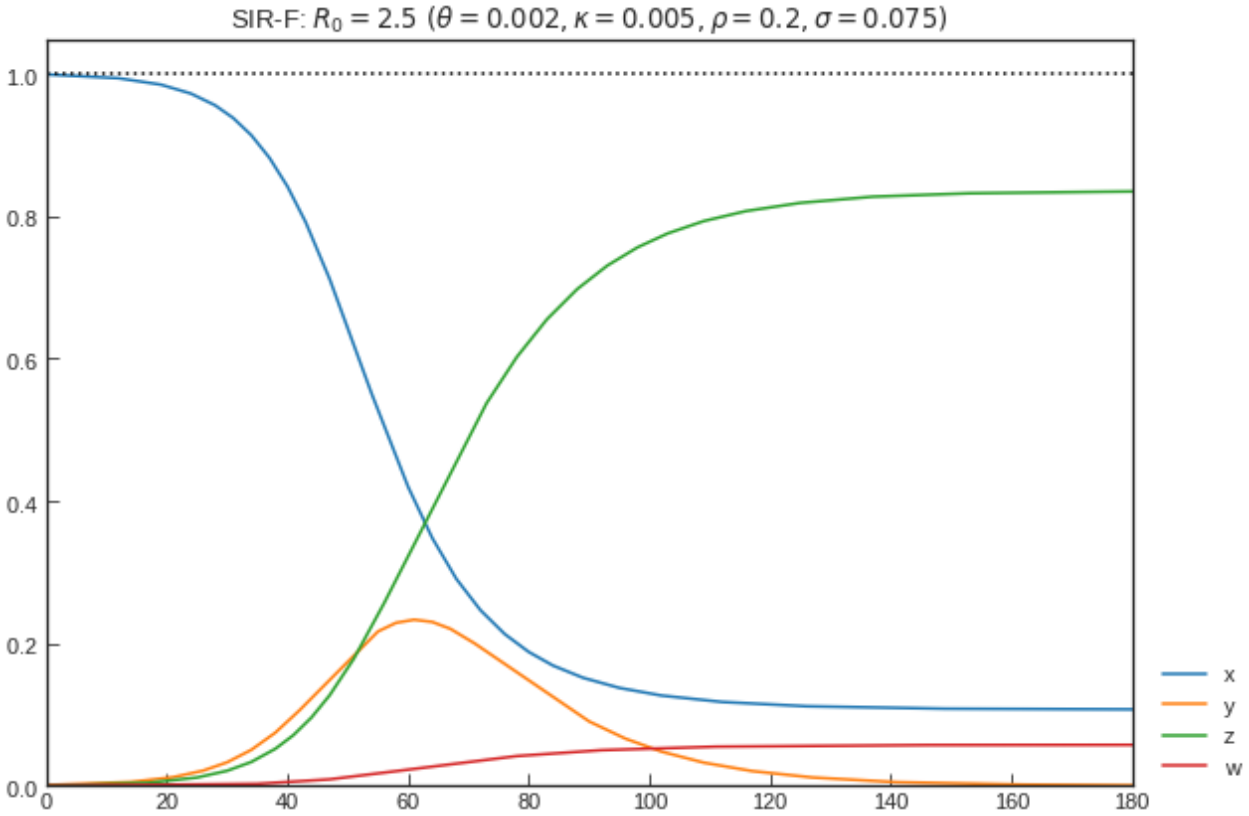


Рисунок 2.6 – Графічне представлення моделі SIR-F [8]

2.5 Висновок до розділу 2

В даному розділі ми розглянули епідеміологічні моделі для аналізу протікання різнотипних захворювань, вказали їх математичне представлення

та привели приклади їхнього загального вигляду. На основі розглянутого матеріалу було обрано найкращий варіант з представлених моделей для опрацювання обраної теми та аналізу її результатів.

РОЗДІЛ 3 АРХІТЕКТУРА ТА АНАЛІЗ ПРОГРАМИ

3.1 Використані програмні засоби

Розробка програмного модуля відбувалася на мові програмування Python 3.8. Цей вибір був зроблений, на основі основних переваг даної мови програмування, таких як: велика функціональність – велика кількість вже розроблених модулів, актуальність – на сьогоднішній день є однією з найпопулярніших та найрозповсюдженіших мов програмування, та багатоплатформність – тобто можливість в виконанні програмного продукту на базі персонального комп’ютера або ж, в якості веб-додатку.

В ході розробки програми були використані такі модулі як pandas, numpy, matplotlib, за допомогою яких відбувалась обробка, підготовка та візуалізація даних.

3.2 Підготовка даних

Для опрацювання даної роботи використовувались дані про міграційні потоки населення Сполучених Штатів Америки, з національного порталу Census [17]. Інформація про міграційні потоки була представлена на рівні державних округів, але через низькі середні показники було прийнято рішення опрацьовувати їх на рівні штатів. З загальним виглядим інформації про міграцію можна ознайомитись на Рисунку 3.1.

	FIPS	State	County_Name	Estimate	MOE	POPESTIMATE2019
0	01001	Alabama	Autauga County	4367.0	6720	55869.0
1	01003	Alabama	Baldwin County	10505.0	14864	223234.0
2	01005	Alabama	Barbour County	1227.0	1817	24686.0
3	01007	Alabama	Bibb County	1712.0	1898	22394.0
4	01009	Alabama	Blount County	1975.0	3056	57826.0
...	...	...	...	...	...	...

Рисунок 3.1 Загальний вигляд вхідних даних [9]

State	Estimate
Texas	1621404.0
California	1580553.0
Florida	1209244.0
Georgia	750873.0
New York	748434.0
Virginia	678138.0
North Carolina	659778.0

Рисунок 3.2 Рейтинг штатів за показником міграції [10]

З огляду на Рисунок 3.1 визначимо стовпці: FIPS – кодовий номер для кожного округу чи штату, визначений за державним стандартом США; State – штат; County_Name – округ відповідного штату; Estimate – оцінка міграції в відповідному окрузі; MOE – похибка оцінки; POPESTIMATE2019 – загальна кількість населення округу на 2019 рік [18]. На Рисунку 3.2 можемо побачити рейтинг штатів з найвищими показниками населення.

Для оцінки швидкості розповсюдження коронавірусу було прийнято рішення обрати для аналізу інформацію з відомого порталу Johns Hopkins University [19-20]. Після приведення даних до потрібного вигляду вони мають вигляд відповідно до Рисунку 3.3.

	Date	State	Province	Confirmed	Infected	Fatal	Recovered
0	2020-01-22	Washington	-	2	2	0	0
1	2020-01-23	Washington	-	2	2	0	0
2	2020-01-24	Washington	-	2	2	0	0
3	2020-01-24	Illinois	-	1	1	0	0
4	2020-01-25	Washington	-	2	2	0	0
...	...	...	...	...	...	...	...
4603	2020-06-01	Nebraska	-	14130	13953	177	0
4604	2020-06-01	Nevada	-	8634	7198	422	1014
4605	2020-06-01	Arizona	-	20123	20123	0	0
4606	2020-06-01	New Jersey	-	160055	148334	11721	0
4607	2020-06-01	Wyoming	-	910	910	0	0

Рисунок 3.3 Дані про протікання вірусу [11]

Як зазначалось раніше, було прийнято рішення про аналіз на рівні штатів, тому на Рисунку 3.3 ми можемо спостерігати за розвитком показників зафіксованих випадків – Confirmed, інфікованих - Infected, кількість випадків, коли вдалося вилікуватись – Recovered, кількість летальних випадків – Fatal, відповідно за штатом – State та датою – Date. Загальна розвиток епідемії на всій території Сполучених Штатів Америки можна спостерігати на Рисунку 3.4.

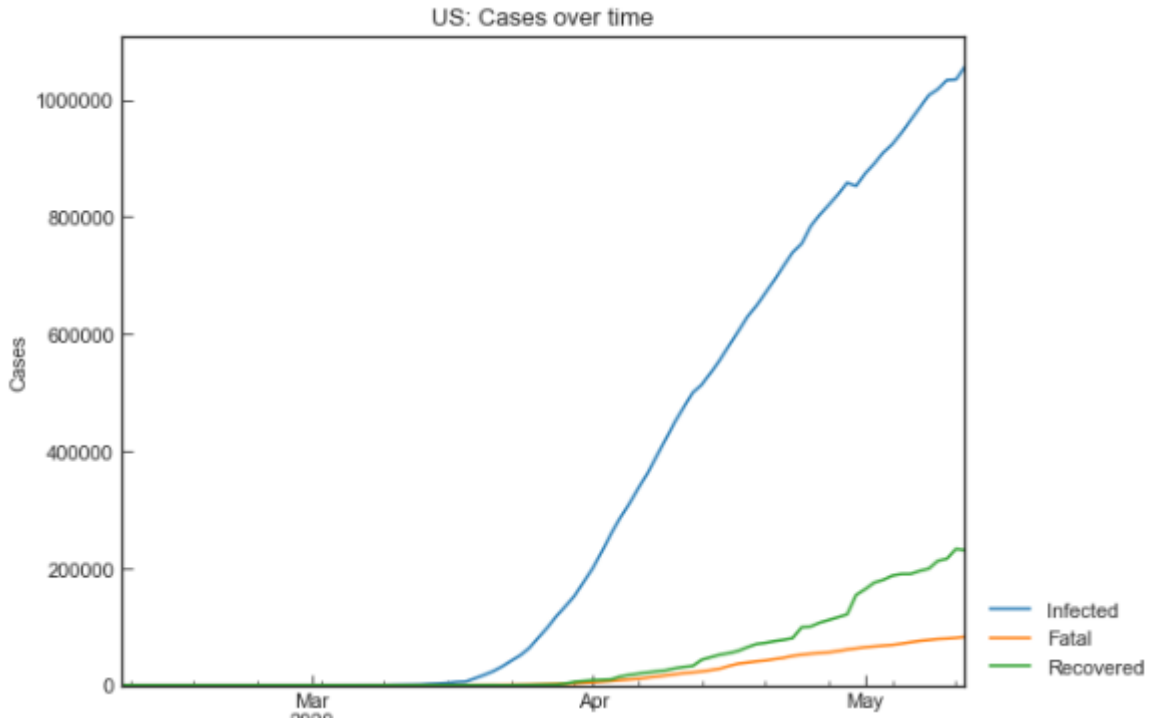


Рисунок 3.4 [12]

3.3 Моделювання епідемії

Щоб проаналізувати швидкість розповсюдження коронавірусу спочатку потрібно створити модель для аналізу показників. Для моделювання протікання епідемії була обрана модель SIR-F, якою можна ознайомитись в Розділі 2.

На вхід ми подаємо історичні дані з епідеміологічними показниками та показники популяції для кожного штату (Рисунок 3.1). На основі цих даних створюємо графічну анімацію для зображення розповсюдження епідемії по всій країні (Рисунок 3.5-3.7).

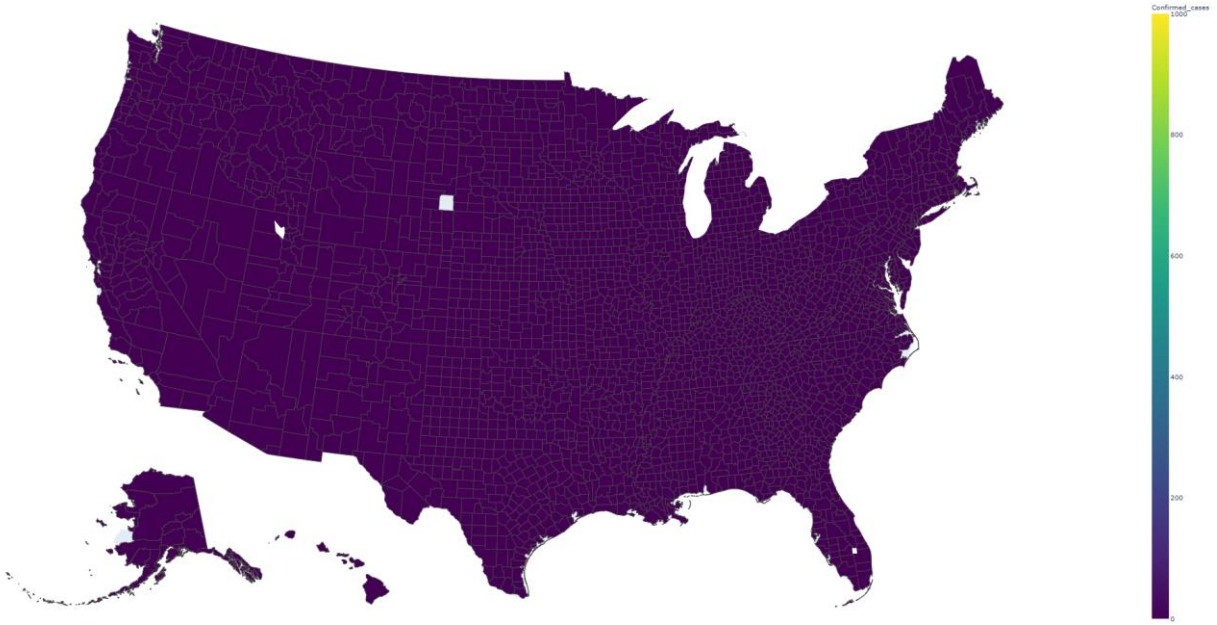


Рисунок 3.5 Початок епідемії [13]

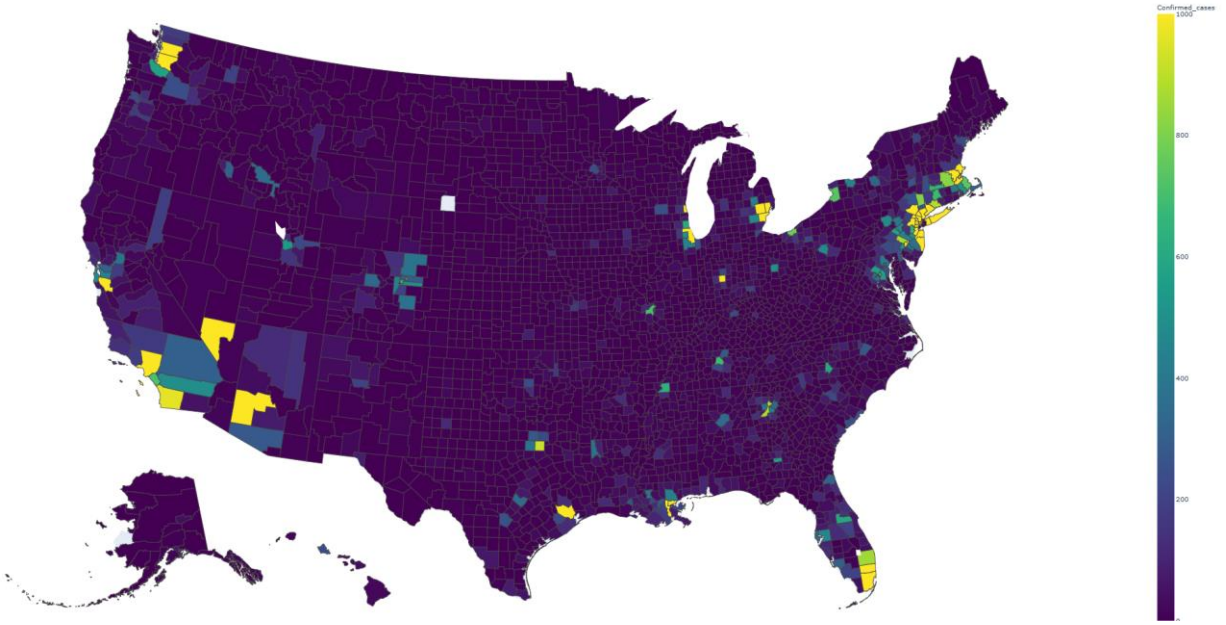


Рисунок 3.6 Розповсюдження вірусу по території [14]

Через велику градацію випадків по штатах було прийнято рішення про прийняття верхньої границі в 1000 підтверджених випадків, для зручності подання інформації.

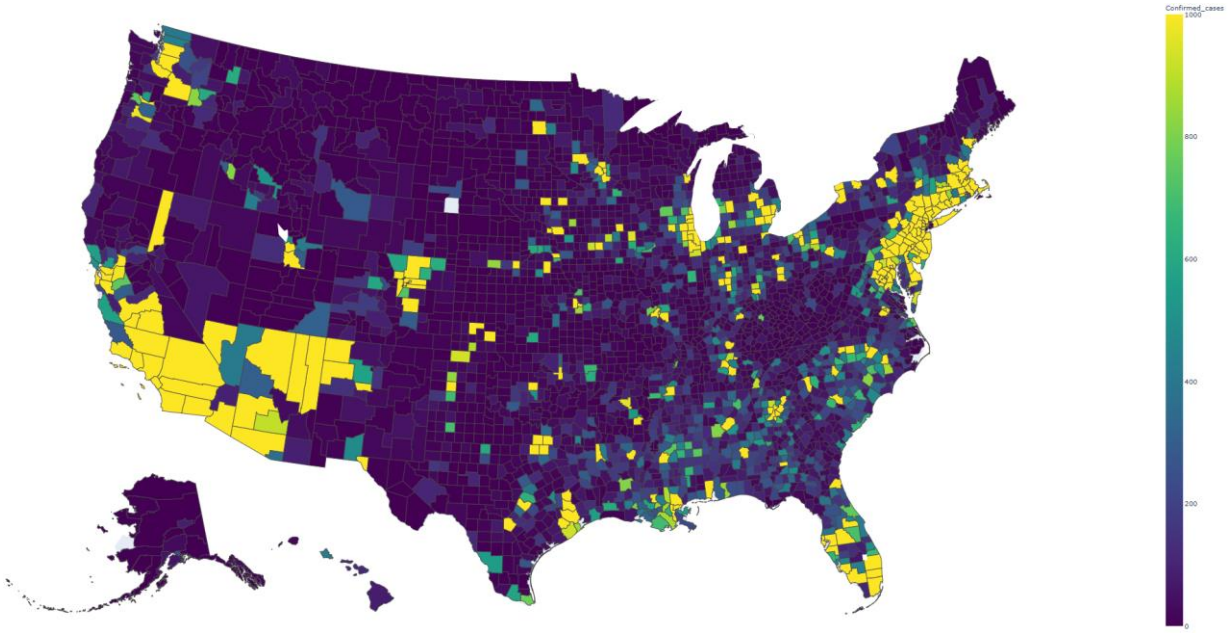


Рисунок 3.7 Епідеміологічний стан на момент виконання роботи [15]

Одразу можемо порівняти нинішній рівень випадків захворювання з рівнем міграції по штатам (Рисунок 3.8).

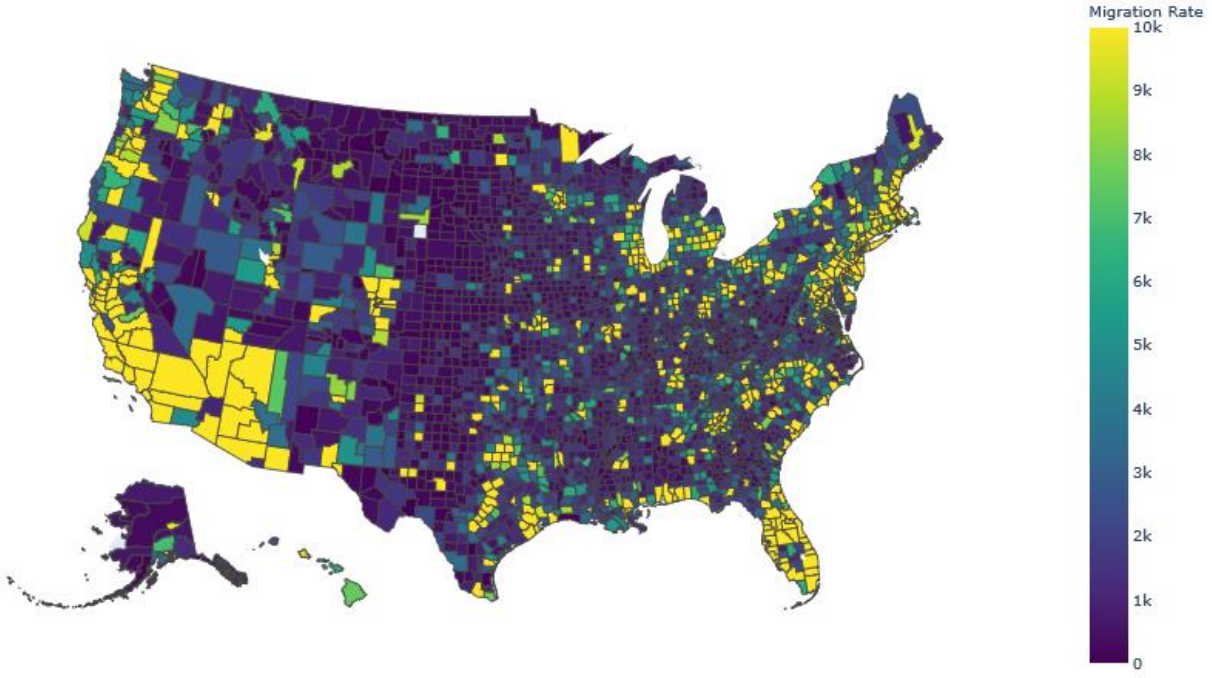


Рисунок 3.8 Рівень міграції по штатам [16]

Для оцінки майбутніх показників нам потрібно проаналізувати як змінюються параметри на історичних даних та виділити певні тренди, коли вони були схожі між собою. Кількість поділів на тренди задаються самостійно, тому проаналізувавши ситуацію в усіх штатах було прийнято обрати кількість трендів n_trend = 3. Розглянемо на прикладі штату Каліфорнія як відбувається тренд аналіз (Рисунок 3.9-3.11).

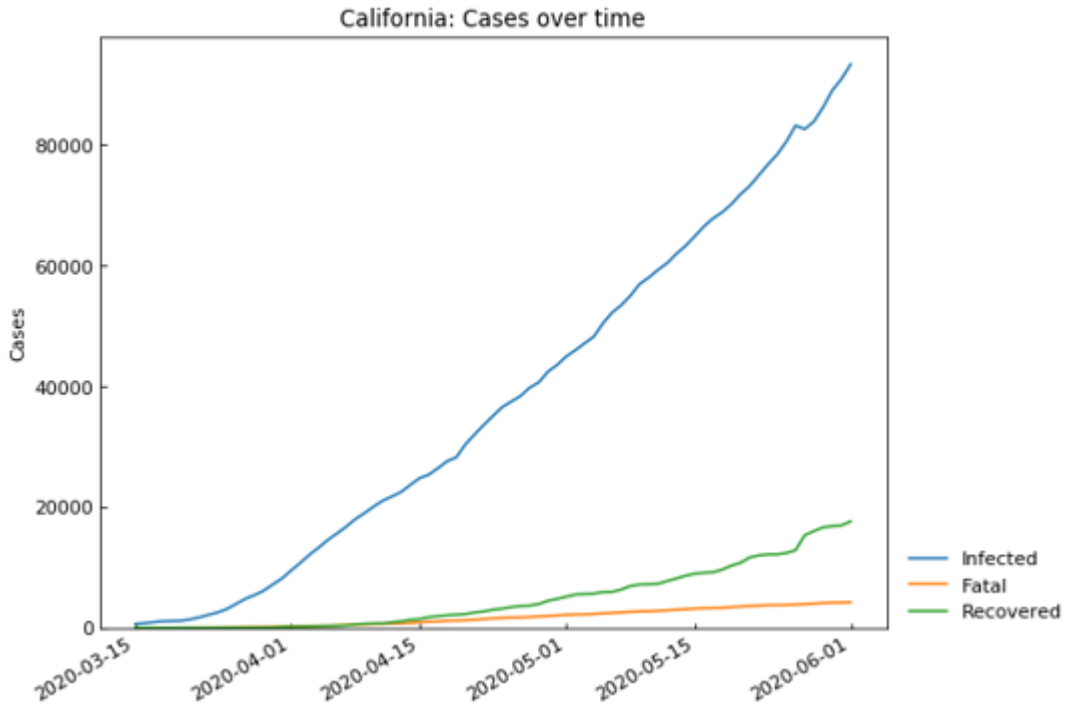


Рисунок 3.9 Загальне протікання хвороби для штату Каліфорнія [17]

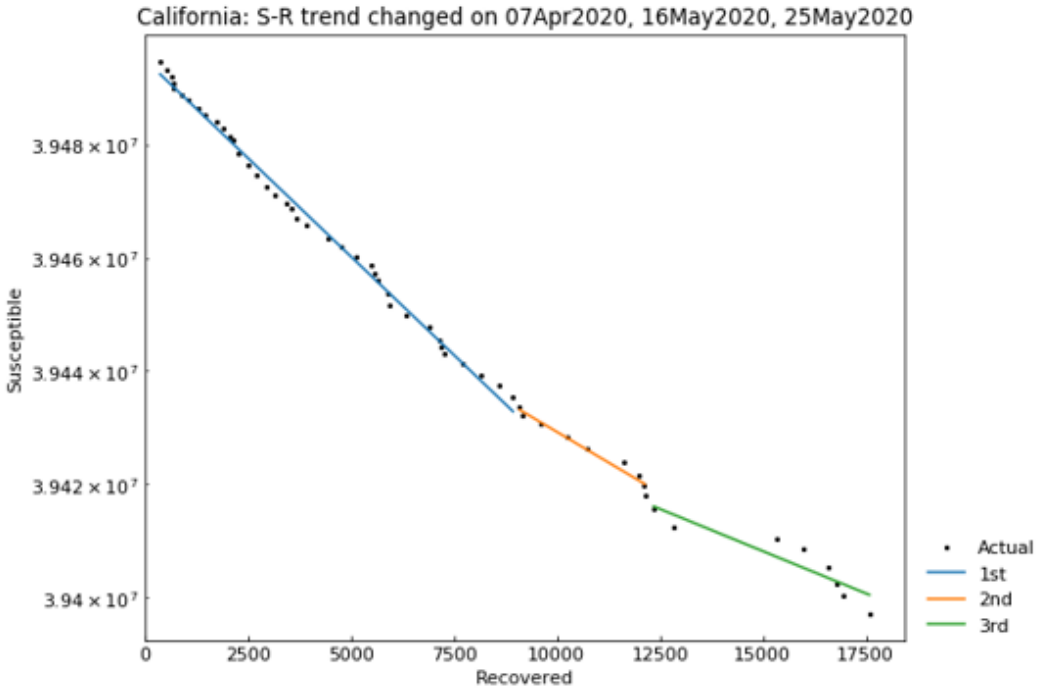


Рисунок 3.10 Поділ на тренди за схожими показниками [18]

На Рисунку 3.10 ми спостерігаємо, як відбувається поділ на тренди, виділивши схожу динаміку. На Рисунку 3.11 зображені епідеміологічні параметри моделі з відповідними періодами для штату Каліфорнія, які надалі будуть необхідні нам для аналізу швидкості розповсюдження коронавірусу по всій території країни.

	Type	Start	End	Population	ODE	tau	theta	kappa	rho	sigma	Rt	alpha1 [-]	1/alpha2 [day]	1/beta [day]	1/gamma [day]	RMSLE
1st	Past	06Apr2020	10May2020	39512223	SIR-F	111	0.025055	0.000052	0.003728	0.000455	7.17	0.025	1470	20	169	0.336508
2nd	Past	11May2020	22May2020	39512223	SIR-F	111	0.050672	0.000046	0.002293	0.000428	4.59	0.051	1684	33	180	0.0768706
3rd	Past	23May2020	01Jun2020	39512223	SIR-F	111	0.038015	0.000042	0.002209	0.000419	4.60	0.038	1831	34	183	0.0872706
4th	Future	02Jun2020	09Jun2020	39512223	SIR-F	111	0.038015	0.000042	0.002209	0.000419	4.60	0.038	1831	34	183	-

Рисунок 3.11 Епідеміологічні показники моделі [19]

Окрім раніше визначених параметрів ми Рисунку 3.11 є параметр RMSLE, який відповідає похибці моделі, та обчислюється за формулою:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log_{10}(A_i + 1) - \log_{10}(P_i + 1))^2}$$

де A та P з відповідними індексами відповідають історичним та прогнозованим значенням S, I, R, F.

3.4 Оцінка параметрів моделі та зв'язок між ними

Отримавши сценарій протікання хвороби та епідеміологічні показники, можемо розглянути зв'язок між розповсюдженням коронавірусу та міграційними потоками. На Рисунку 3.12 розглянемо основну тенденцію по штатам між підтвердженими випадками та міграцією в штаті.

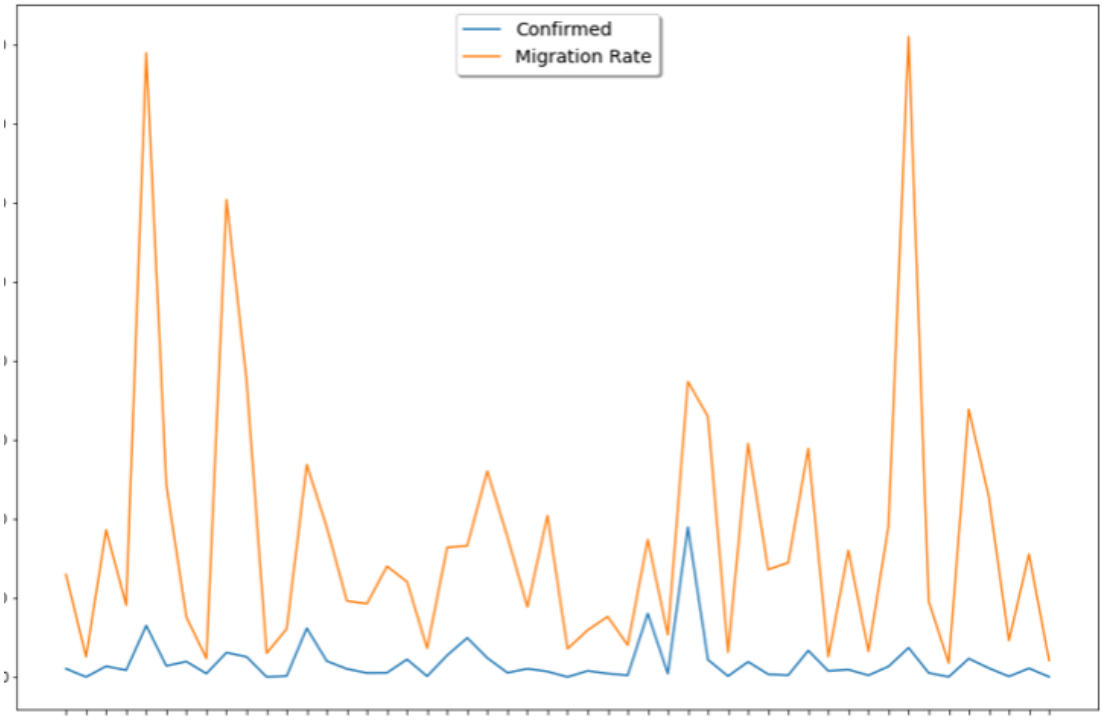


Рисунок 3.12 Показник міграції та підтверджені випадки захворювання [20]

Як видно з Рисунку 3.12 по графіку не на всіх штатах можна помітити тенденцію залежності. Такий результат може пояснюватись похибкою в моделі, або ж неточністю початкових даних, частина з яких була відсутня. Розглянемо які штати лідирують по показникам швидкості захворювання та рейтингу міграції (Рисунок 3.13).

	State	1/beta [day]
0	Idaho	66
1	West Virginia	63
2	Florida	52
3	Wyoming	51
4	Washington	47
5	Tennessee	42
6	South Dakota	40
7	Connecticut	37
8	Oregon	37
9	New York	34

	State	Estimate
39	Texas	1621404
4	California	1580553
8	Florida	1209244
9	Georgia	750873
28	New York	748434
42	Virginia	678138
29	North Carolina	659778
31	Ohio	591360
34	Pennsylvania	579098
12	Illinois	538151

Рисунок 3.13 Штати за найбільшими показниками міграції та швидкості розповсюдження вірусів [21]

З огляду на ці показники можемо зробити висновок, що прямої залежності між показниками міграції та швидкості розповсюдження коронавірусу немає. Але слід прийняти до уваги і фактор густоти населення, який відіграє важливу роль в цьому процесі. Так – в штаті Техас густота населення є дуже низькою, а це негативно впливає на швидкість розповсюдження, але, наприклад, штат Каліфорнія є дуже густонаселеним, тому і показник швидкості розповсюдження є високим, що і видно на Рисунку 3.13

Розглянемо як між собою корелюють епідеміологічні показники та рейтинг міграції в цілому (Рисунок 3.14).

	__Estimate	Confirmed	Fatal	Recovered	Infected
__Estimate	1.000000	0.657143	0.564082	0.071336	0.550879
Confirmed	0.657143	1.000000	0.766531	0.131762	0.691459
Fatal	0.564082	0.766531	1.000000	0.083086	0.568860
Recovered	0.071336	0.131762	0.083086	1.000000	-0.038653
Infected	0.550879	0.691459	0.568860	-0.038653	1.000000

Рисунок 3.14 Кореляція показників [22]

Як бачимо кількість випадків досить непогано корелює з міграційним рейтингом. Проте такий середній показник, може свідчити проте, що залежність між даними не досить сильна, що може пояснюватись різною щільністю населення, відсутністю коректних даних, або ж навіть швидкістю життя по штатах.

3.5 Висновок до розділу 3

В ході даного розділу ми розглянули етапи та процес розробки програмного продукту. Проаналізувавши отримані результати ми зрозуміли, що для дослідження таких масштабних процесів, як спалахи епідемій потрібно мати велику вибірку різноманітних даних, для точної оцінки параметрів та подальшого їх аналізу. Ми показали, що між внутрішніми потоками населення США та розповсюдження захворювання COVID-19 є певний зв’язок, хоча і він і не є чітким для виявлення прямої залежності в швидкості поширення (Рисунки 3.7, 3.8,3.12).

РОЗДІЛ 4 ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ ПРОГРАМНОГО ПРОДУКТУ ДЛЯ ВИЗНАЧЕННЯ ЗВ'ЯЗКУ МІЖ ШВІДКІСТЮ РОЗПОВСЮДЖЕННЯ COVID-19 ТА РЕГІОНАЛЬНИМИ МІГРАЦІЙНИМИ ПОТОКАМИ НА ПРИКЛАДІ США

4.1 Постановка задачі проектування

У даному розділі проводиться оцінка програмного продукту призначеного для визначення зв’язку між швидкістю поширення вірусу nCov-2019 та міграційними потоками в США. Програма була створена на мові програмування Python.

Середовищем розробки є Jupyter Notebook що дозволяє надалі вбудовувати даний модуль до веб-застосунків. Нижче наведено аналіз різних варіантів реалізації модуля з метою вибору оптимального з огляду при цьому на основні фактори які впливають на роботу програми та якість вихідного продукту.

4.2 Обґрунтування функцій програмного продукту

Основні функції продукту:

1. F_1 – масштаб побудови моделі.
2. F_2 – алгоритм, що використовується під час досліджень.
3. F_3 – вибір мови програмування.

Функція F_1:

а) дослідження на рівні штатів;

б) дослідження на рівні округів;

Функція F_2:

а) використання існуючого алгоритму;

б) використання алгоритму створеного власноруч;

Функція F_3:

а) мова програмування C++;

б) мова програмування Python;

Побудуємо морфологічну карту за даними варіантами.

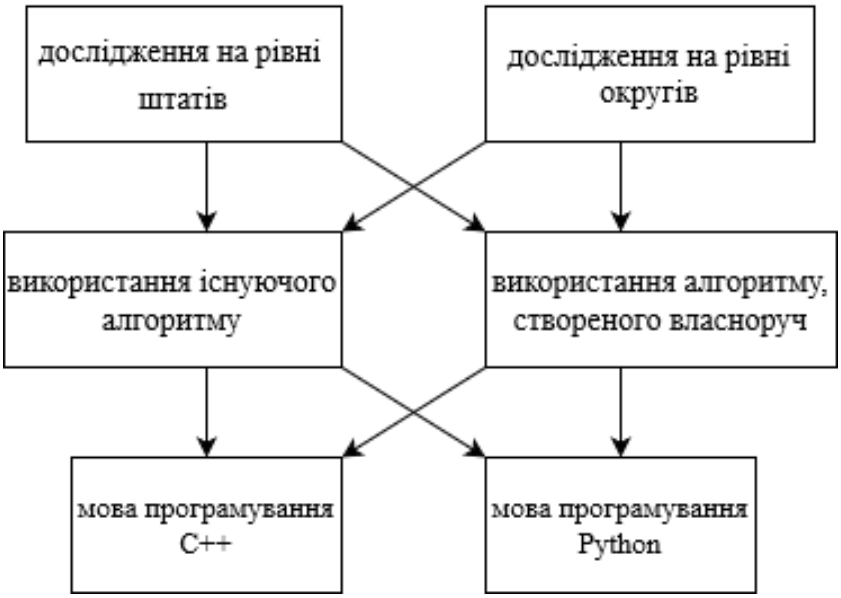


Рисунок 4.1 Морфологічна карта [23]

Спираючись на морфологічну карту побудуємо позитивно-негативну матрицю (Таблиця 4.1).

Таблиця 4.1 Позитивно-негативна матриця

Основні функції	Варіанти реалізації	Переваги	Недоліки
F1	А	Швидкість в реалізації та виконанні продукту	Неточність результатів
	Б	Можливість враховувати внутрішні динамічні фактори	Складність та довга тривалість реалізації
F2	А	Перевірений та відкалібрований алгоритм	Не унікальність алгоритму
	Б	Створення цілеспрямованого алгоритму під необхідні цілі	Необхідність в створенні власного алгоритму
F3	А	Висока швидкість виконання	Великі втрати часу в написанні гнучкої моделі
	Б	Великий функціонал доступних модулів	Порівняно низька продуктивність

Провівши аналіз позитивно-негативної матриці, складемо найбільш оптимальні варіанти:

$$F_1\text{Б} - F_2\text{А} - F_3\text{Б}$$
$$F_1\text{Б} - F_2\text{Б} - F_3\text{Б}$$

Для оцінювання якості моделі використаємо параметри, описані нижче. Визначимо мінімальні, середні та максимальні значення(Таблиця 4.2).

Таблиця 4.2 Система параметрів додатку

Найменування параметрів	Позначення параметру	Значення параметру		
		Мінімальне	Середнє	Максимальне
Час на побудову модуля (год)	X1	12	20	40
Час на оволодіння теорією (год)	X2	20	45	80
Час на калібрування моделі (год)	X3	2	8	20

Відповідно до значень в таблиці представимо графічне зображення параметрів.

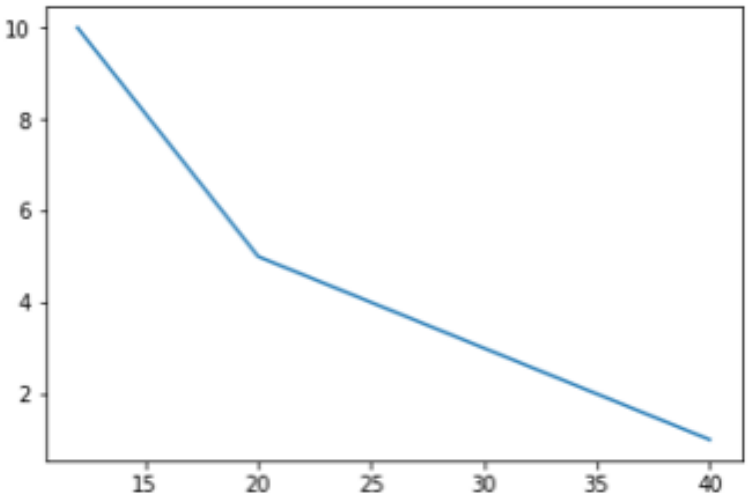


Рисунок 4.2 – Значення параметру X1 [24]

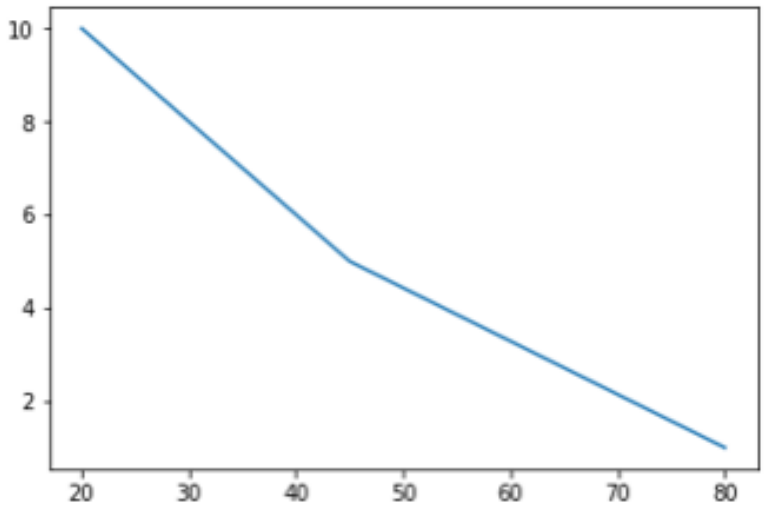


Рисунок 4.3 – Значення параметру X2 [25]

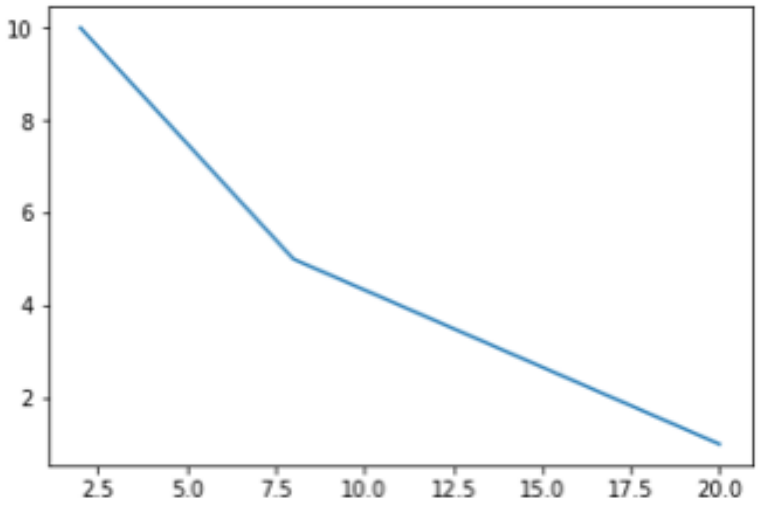


Рисунок 4.4. – Значення параметру X3 [26]

Вагомість параметрів оцінюється за допомогою методів попарного зрівняння. Ранги варіюються від 1 до 5. Результати наведені в Таблиці 4.3-4.4.

Таблиця 4.3 Результат оцінки параметрів

Параметри	Ранг параметру по оцінці експерта							Сума рангів, R_i	Відхилення Δ_i	Квадрат відхилення, $(\Delta_i)^2$
	1	2	3	4	5	6	7			

X1	1	2	2	2	2	1	2	12	-2	4
X2	2	1	1	1	1	2	1	9	-5	25
X3	3	3	3	3	3	3	3	21	7	49
Разом	6	6	6	6	6	6	6	42	0	78

Найбільшим вважаємо ранг 3, відповідно найменшим – 1. Проведемо попарне порівняння всіх параметрів (Таблиця 4.4.).

Таблиця 4.4

Параметри	Експерти							Підсумкова оцінка	Коефіцієнт ваги (хij)
	1	2	3	4	5	6	7		
X1 та X2	<	>	>	>	>	<	>	>	1.5
X2 та X3	<	<	<	<	<	<	<	<	0.5
X3 та X1	>	>	>	>	>	>	>	>	1.5

Обчислимо коефіцієнт координації:

$$W = \frac{12S}{N^2(n^3 - n)} = \frac{12 * 78}{7^2(3^3 - 3)} = 0.79 > W_k = 0.67$$

Отже результат вважаємо достовірним. За результатами попарного порівняння обчислюємо вагомість кожного з критеріїв (Таблиця 4.5).

Таблиця 4.5

Параметри	Параметри			Перший крок		Другий крок	
	X1	X2	X3	b_i	$К_{вi}$	$b_i{}^1$	$\hat{E}_{\hat{a}^3}{}^1$
X1	1	1.5	0.5	3	0,33	8	0.32

X2	0.5	1	0.5	2	0,22	5.5	0.22
X3	1.5	1.5	1	4	0,44	11.5	0.46
Загалом:				9	1	25	1

Як видно з таблиці різниця значень коефіцієнтів вагомості не перевищує 2%, тому більшої кількості ітерацій не потрібно.

4.3 Аналіз варіантів реалізації функцій

Таблиця 4.6

Основні функції	Варіант реалізації функції	Параметри	Абсолютне значення параметра	Бальна оцінка параметра	Коефіцієнт вагомості параметра	Коефіцієнт рівня якості
F1	Б	X1	20	5	0.32	1.6
		X2	40	6	0.22	1.32
F2	А	X1	12	10	0.32	3.2
		X2	35	7	0.22	1.54
	Б	X1	15	8	0.32	2.56
		X2	40	6	0.22	1.32
F3	Б	X3	5	7.33	0.46	3.37

Обчислимо рівень якості для наших варіантів:

$$К_1 = 1.6 + 1.32 + 3.2 + 1.54 + 3.37 = 11.03$$

$$К_2 = 1.6 + 1.32 + 2.56 + 1.32 + 3.37 = 10.17$$

Як бачимо, перший варіант виявився кращим.

4.4 Економічний аналіз варіантів розробки

Для оцінки вартості розробки проведемо розрахунок трудомісткості. Обидва варіанти мають такі основні завдання:

1. Розробка алгоритму.
2. Розробка програмного продукту.

Окрім цього кожен з варіантів має додаткове завдання (3.1. – для першого, 4.1 – для другого).

3.1 Освоєння теоретичної бази для роботи з алгоритмом.

4.1 Обробка даних для створення моделі.

За складністю алгоритмів до групи 1 відносять завдання 1 та 2, до групи складності 3 – 3.1 та 4.1. За степенем новизни до групи А належить завдання 2 та 4.1, до групи В - 1 та 3.1. Спираючись на норми розрахункового часу визначимо трудомісткість.

Для завдання 1: $Т_р = 43$ людино-дні, $К_п = 0.81, К_{ст} = 0.7,\ К_{ск} = 1,$ $К_{ст.м} = 1$. Звідси маємо, що Т1 = 43 * 0.7 * 0.81 = 24.383 людино дні.

Для завдання 2: $Т_р = 90$ людино-дні, $К_п = 1.7, К_{ск} = 1, К_{ст.м} = 1,$ $К_{ст} = 0.7$. Звідси маємо, що Т2 = 90 * 0.7 * 1.7 = 107.1 людино дні.

Для додаткового завдання 3.1: $Т_р = 12$ людино-дні, $К_п = 0.6, К_{ск} = 1,$ $К_{ст.м} = 1, К_{ст} = 0.7$. Звідси маємо, що Т3 = 12 * 0.6 * 0.7 = 5.04 людино дні.

Для додаткового завдання 4.1: $Т_р = 27$ людино-дні, $К_п = 1.26, К_{ск} = 1,$ $К_{ст.м} = 1, К_{ст} = 0.7$. Звідси маємо, що Т4 = 27 * 0.7 * 1.26 = 23.814 людино дні. Отже повна трудомісткість варіантів:

$$T_1 = 24.383 + 107.1 + 5.04 = 136.523$$
$$T_2 = 24.383 + 107.1 + 23.814 = 155.29$$

Другий варіант є більш трудомістким.

В розробці бере участь один програміст, з окладом 15000 та тестувальник з окладом 12000. Визначимо погодинну заробітну плату:

$$C_ч = \frac{14000 + 12000}{2 * 21 * 8} = 77.38,$$

Відповідно за варіантами заробітна плата:

1) $C_{ЗП} = 77.38 * 8 * 136.523 = 84513.1979$
2) $C_{ЗП} = 77.38 * 8 * 155.297 = 96135.0549$

Відрахування ЄСВ 22%:

1) $C_{ВІД} = 84513.1979 * 0.22 = 18592.9035$
2) $C_{ВІД} = 96135.0549 * 0.22 = 21149.7121$

Визначаємо витрати на оплату однієї машино-години. З урахуванням заробітної плати програміста в розмірі 15000 грн з коефіцієнтом зайнятості 0.3, отримуємо: $C_г = 12 * 15000 * 0{,}6 = 54000$ грн.

Враховуємо додаткову заробітну плату: $C_{зп} = 54000 * (1 + 0.3) = 70200$ грн.

Відрахування на соціальне страхування: $C_{ВІД} = 70200 * 0.22 = 15444$ грн.

Розрахуємо амортизаційні відрахування при вартості ЕОМ 25000 грн. та амортизації 25%: $C_А = К_{ТМ} * К_А * Ц_{ПР} = 1.15 * 0.25 * 25000 = 7187.5$ грн.
Розрахуємо профілактичні витрати:
$C_р = К_{ТМ} * К_Р * Ц_{ПР} = 1.15 * 0.05 * 25000 = 1437.5$ грн.

Розрахуємо ефективний годинний фонд часу ПК за рік

$$T_{ЕФ} = (365 - 104 - 11 - 16) * 8 * 0{,}8 = 1497.6 \text{ год}$$

Розрахуємо витрати на електроенергію з урахуванням ПДВ:

$$C_{ЕЛ} = 1497.6 * 0.3 * 1.378 * 1.75 = 1083.43 \text{ грн}$$

Накладні витрати: $C_Н = 54000 * 0.67 = 36180$ грн.

Звідси річні експлуатаційні витрати складають:

$$C_{ЕКС} = 70200 + 15444 + 7187.5 + 1437.5 + 1083.43 + 36180 = 131526.43$$

Отже собівартість однієї машино-години ЕОМ дорівнює:

$$C_{М-Г} = \frac{131526.43}{1497.6} = 87.82 \text{ грн/год}$$

Отже, обрахуємо витрати на оплату машинного часу:

1) $C_м = 87.82 * 8 * 136.523 = 95915.5989$
2) $C_м = 87.82 * 8 * 155.297 = 109105.46$

Відповідно накладні витрати:

1) $C_Н = 95915.5989 * 0{,}67 = 64263.4513$
2) $C_Н = 109105.46 * 0{,}67 = 73100.6584$

Повна вартість розробки за варіантами:

1) $C_{ПП} = 84513.2 + 18592.9 + 95915.6 + 64263.4 = 263285.1$
2) $C_{ПП} = 96135.05 + 21149.7 + 109105.46 + 73100.6 = 299490.8$

4.5 Вибір кращого варіанту ПП техніко-економічного рівня.

Розрахуємо коефіцієнт техніко-економічного рівня

$$К_{ТЕР1} = \frac{11.03}{263285.1} = 4.18 * 10^{-5}$$

$$К_{ТЕР2} = \frac{10.17}{299490.8} = 3.39 * 10^{-5}$$

4.6 Висновок до розділу 4

Отже враховуючи всі дослідження, що описані вище, можна сказати, що перший варіант реалізації є найбільш оптимальним з якісно-економічної оцінки. Його коефіцієнт техніко-економічного рівня складає $4{,}18 * 10^{-5}$.

ВИСНОВКИ

В ході даної роботи було створено програмний продукт для аналізу та симуляції епідеміологічних процесів захворювання COVID-19 та знаходження зв’язку між ними та міграційними потоками на внутрішньому рівні Сполучених Штатів Америки.

В процесі дослідження епідеміологічної сфери були розглянути основні моделі для передбачення, аналізу та симуляції різнотипних захворювань та вірусів, починаючи від самих перших спроб створення моделей, до найбільш популярних та найбільш використовуваних алгоритмів дослідження.

На основі проведеного аналізу існуючих моделей та за обраною темою дослідження був створений програмний додаток для аналіз показників захворювання COVID-19 та аналізу впливу міграційних потоків на процес та швидкість розповсюдження захворювання на прикладі Сполучених Штатів Америки.

ПЕРЕЛІК ПОСИЛАНЬ

1. Burkom H.S., Murphy S.P., Shmueli G. Automated Time Series Forecasting for Biosurveillance. 2007.
2. Pelat C., Bo¨elle P.Y., Cowling B.J., Carrat F., Flahault A., Ansart S., Valleron A.J. Online detection and quantification of epidemics. 2007.
3. Serfling R.E. Methods for Current Statistical Analysis of Excess Pneumonia-influenza Deaths. 1963.
4. Sebastiani P. A Bayesian dynamic model for influenza surveillance. 2006.
5. Siettos C.I., Russo L. Mathematical modeling of infectious disease dynamics. 2013.
6. Chatfield C., Yar M. Holt-Winters Forecasting: Some Practical Issues. 1988.
7. Shaman J., Karspeck A. Forecasting seasonal outbreaks of influenza. 2012.
8. Kemac W.O. A Contribution to the Mathematical Theory of Epidemics. 2005.
9. Hethcote, H.W., Van Ark. J.W. Epidemiological models for heterogeneous populations; proportionate mixing, parameter estimation and immunization programs. 2007.
10. Diekmann, O. and H. Heesterbeek. Mathematical Epidemiology of Infectious Diseases. 2000.
11. Hefferman, J., R. Smith, and L. Wahl. Perspectives on the basic reproduction ratio. 2005.
12. Herbert W Hethcote. The basic epidemiology models: Models, expressions for r0, parameter estimation, and applications. 2009.
13. Hethcote, H.W., Van Ark. J.W. Epidemiological models for heterogeneous populations; proportionate mixing, parameter estimation and immunization programs. 1987.

14. F. Berezovsky, G. Karev, B. Song, and C. Castillo-Chavez. “A simple epidemic model with surprising dynamics”. 2005.
15. Dynamics of an SEIR epidemic model with nonlinear incidence and treatment rates. 2019. URL: https://link.springer.com/article/10.1007/s11071-019-04926-6 (дата звернення: 18.05.2020).
16. C Connell McCluskey. Complete global stability for an sir epidemic model with delay—distributed or discrete. 2010.
17. US Statte-to-State Migration Flows. URL: https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-to-state-migration.html (дата звернення: 25.05.2020).
18. US Population Estimates. URL: https://www.census.gov/topics/population.html (дата звернення: 25.05.2020)
19. Johns Hopkins University. URL: https://coronavirus.jhu.edu/data (дата звернення 23.05.2020).
20. John Hopkins Medecine URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus (дата звернення: 23.05.2020)
21. Gevorg Yeghikyan Modelking coronavirus epidemic in a city. URL: https://towardsdatascience.com/modelling-the-coronavirus-epidemic-spreading-in-a-city-with-python-babd14d82fa2 (дата звернення: 20.05.2020)

ДОДАТОК А ТЕКСТ ПРОГРАМИ

migration_data.ipynb

```
import pandas as pd
xls = pd.ExcelFile("county-to-county-2013-2017-ins-outs-nets-gross.xlsx")
sheet_to_df_map = {}
for sheet_name in xls.sheet_names:
    sheet_to_df_map[sheet_name] = xls.parse(sheet_name)
for sheet_name in xls.sheet_names:
    sheet_to_df_map[sheet_name] = sheet_to_df_map[sheet_name].iloc[2:-5,
:10].drop(['Unnamed: 2', 'Unnamed: 3'], axis=1).rename(columns = {
    'Table with column headers in rows 2 through 3.':'State_FIPS',
    'Unnamed: 1':'County_FIPS',
                                    'Unnamed: 4':'State_Name_of_Geography_A',
                                    'Unnamed: 5':'County_Name_of_Geography_A',
                                    'Unnamed: 6':'State_Name_of_Geography_B',
                                    'Unnamed: 7':'County_Name_of_Geography_B',
                                    'Unnamed:
8':'Flow_from_Geography_B_to_Geography_A__Estimate',
                                    'Unnamed: 9':'MOE'})
for sheet_name in xls.sheet_names:
    sheet_to_df_map[sheet_name]["FIPS"] =
sheet_to_df_map[sheet_name]["State_FIPS"] +
sheet_to_df_map[sheet_name]["County_FIPS"]
    sheet_to_df_map[sheet_name].drop(['State_FIPS', 'County_FIPS'], axis=1,
inplace=True)
for sheet_name in xls.sheet_names:
```

```
    sheet_to_df_map[sheet_name] = sheet_to_df_map[sheet_name][~sheet_to_df_map[sheet_name].isin(['-'])].dropna()
    for sheet_name in xls.sheet_names:
    sheet_to_df_map[sheet_name] = sheet_to_df_map[sheet_name].groupby(['FIPS', 'County_Name_of_Geography_A'], as_index=False)['Flow_from_Geography_B_to_Geography_A__Estimate', 'MOE'].sum()
for sheet_name in xls.sheet_names:
    sheet_to_df_map[sheet_name]["State"] = sheet_name
    sheet_to_df_map[sheet_name]["FIPS"] = sheet_to_df_map[sheet_name]["FIPS"].map(lambda x: str(x)[1:])
all_counties = pd.DataFrame(columns = ['FIPS',
                                       'County_Name_of_Geography_A',
                                       'Flow_from_Geography_B_to_Geography_A__Estimate',
                                       'MOE',
                                       'State'])
for sheet_name in xls.sheet_names:
    all_counties = pd.concat([all_counties, sheet_to_df_map[sheet_name]], ignore_index=True)
new = pd.DataFrame(all_counties)
one = new.iloc[:-78]
one = one.rename(columns={'County_Name_of_Geography_A':'County_Name', 'Flow_from_Geography_B_to_Geography_A__Estimate':'Estimate'})
one.to_csv('migration_by_county.csv', index=False)
migration = pd.read_csv('migration_by_county.csv', dtype={'FIPS':'str', 'Estimate':'Int64'})
from urllib.request import urlopen
import json
```

```
with urlopen('https://raw.githubusercontent.com/plotly/datasets/master/geojson-counties-fips.json') as response:
    counties = json.load(response)

import plotly.express as px

fig = px.choropleth(migration, geojson=counties,  color='Estimate',
                    locations='FIPS',
                    range_color=(0, 10000),
                    color_continuous_scale="Viridis",
                    scope="usa",
                    labels={'Estimate':'Migration Rate'}
                   )
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":0})
fig.show()

result.ipynb
import pandas as pd
import matplotlib.pyplot as plt
trend = pd.read_csv('trend.csv')
trend['State'].unique()
migration = pd.read_csv('migration_by_county.csv', dtype={'FIPS':'str', 'Estimate':'Int64'})
gr = migration.groupby('State').sum()
trend
f = {'tau':'mean','theta':'mean','kappa':'mean','rho':'mean','sigma':'mean','1/beta [day]':'max', "Rt":'mean'}
t = trend.copy()
t_gr = t.groupby('State').agg(f)
```

```
t_gr.to_csv('t_gr.csv')
tt_gr = pd.read_csv('t_gr.csv')
migr = pd.read_csv('gr.csv')
pr =pd.read_csv('prediction_by_state.csv')
temp = pr[pr['Date']=='06Jun2020']
mm = migr.merge(temp, left_on='State', right_on='State')
mmm = mm[['Estimate', 'Confirmed']]

fig, ax = plt.subplots()
plt.rcParams["figure.figsize"] = (15, 10)
ax.plot(mmm['State'], mmm['Confirmed'], label='Confirmed')
ax.plot(mmm['State'], mmm['Estimate'], label = 'Migration Rate')
legend = ax.legend(loc='upper center', shadow=True, fontsize='x-large')

plt.show()
mm.corr()
```

prediction.ipynb

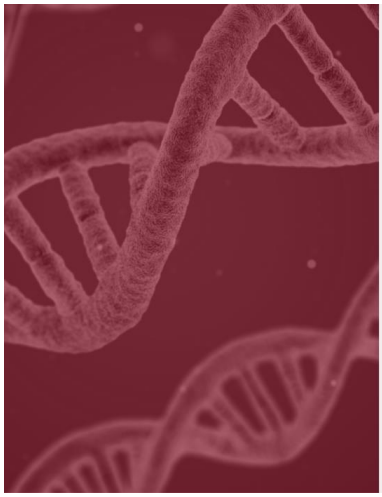
```
from datetime import datetime
time_format = "%d%b%Y %H:%M"
datetime.now().strftime(time_format)
import covsirphy as c
import pandas as pd
pop = c.Population('population_state.csv'
)
# pop.cleaned()
population_dict = pop.to_dict(country_level=True)
# population_dict
jhu_data = c.JHUData("df.csv")
```

```
ncov_df = jhu_data.cleaned()
# print(ncov_df)
list =['Washington', 'Illinois', 'California', 'Arizona', 'Massachusetts',
      'Wisconsin', 'Texas', 'Nebraska', 'Utah', 'Oregon', 'Rhode Island',
      'Florida', 'New Hampshire', 'New York', 'Georgia',
      'North Carolina', 'New Jersey', 'Tennessee', 'Colorado', 'Nevada',
      'Maryland', 'Pennsylvania', 'Oklahoma', 'Hawaii', 'Kentucky',
      'Minnesota', 'South Carolina', 'Indiana', 'Kansas', 'Vermont',
      'Missouri', 'Iowa', 'Connecticut', 'Louisiana', 'Ohio', 'Michigan',
      'South Dakota', 'Wyoming', 'Delaware', 'New Mexico', 'Mississippi',
      'North Dakota', 'Arkansas', 'Alaska', 'Maine', 'Montana', 'Idaho',
      'Alabama', 'West Virginia', 'Virginia']
df = pd.DataFrame(columns=['Date', 'State', 'Confirmed', 'Fatal', 'Infected',
'Recovered'])
for item in list1:

   usa_scenario = c.Scenario(jhu_data, pop, item)
   usa_scenario.records().tail()
   _ = usa_scenario.trend()
   usa_scenario.trend(n_points=3)

   # usa_scenario.summary()

   usa_scenario.estimate(c.SIRF)

   usa_scenario.estimate_accuracy(phase="1st")
   usa_scenario.estimate_accuracy(phase="2nd")
   usa_scenario.estimate_accuracy(phase="3rd")
   y = usa_scenario.summary()
```

```
    y = pd.DataFrame(y)
    y['State'] = item
    est = pd.concat([est, y], ignore_index=True)
    usa_scenario.clear()
    usa_scenario.add_phase(days=7)
    temp = usa_scenario.simulate()
    temp = pd.DataFrame(temp)
    temp['State'] = item
    df = pd.concat([df, temp], ignore_index=True)
    temp = temp[0:0]
migr = pd.read_csv('gr.csv')
migr = migr.sort_values(by=['State'])
migr = migr.drop(migr[migr['State']=='Indiana'].index, axis=0)
migr = migr.drop(migr[migr['State']=='Minnesota'].index, axis=0)
migr = migr.drop(migr[migr['State']=='District of Columbia'].index, axis=0)
param = ['tau', 'theta', 'kappa', 'rho', 'sigma', 'Rt', 'alpha1 [-]',
      '1/alpha2 [day]', '1/beta [day]', '1/gamma [day]', 'RMSLE', 'Trials']
for i in param:
    aa.plot(x='State', y=i)


make_gif.ipynb
import pandas as pd
import matplotlib.pyplot as plt
import glob
import moviepy.editor as mpy
gif_name = 'COVID1.gif'
fps = 5
file_list = glob.glob('png/*')
clip = mpy.ImageSequenceClip(file_list, fps=fps)
```

```
clip.write_gif('{}.gif'.format(gif_name), fps=fps)
```

ДОДАТОК Б ІЛЮСТРАТИВНИЙ МАТЕРІАЛ



Модель зв'язку між швидкістю розповсюдження COVID-19 та регіональними міграційними потоками на прикладі США

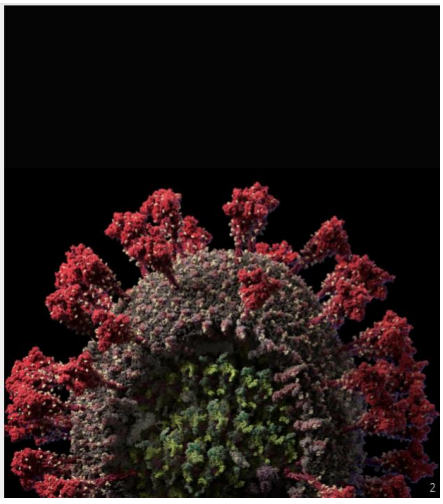
Виконав: студент IV курсу, групи КА-63
Ніколаєнко Богдан

Керівник: Макуха Михайло Павлович

ОБ'ЄКТ ДОСЛІДЖЕННЯ
Регіональні дані про пересування людей між штатами та дані епідеміологічні дані про розвиток епідемії COVID-19.

ПРЕДМЕТ ДОСЛІДЖЕННЯ
Сучасні моделі для аналізу та симуляції епідеміологічних процесів різного типу.

МЕТА РОБОТИ
Розробка програмного продукту для виявлення зв'язку між швидкістю розповсюдження захворювання COVID-19 та внутрішніми міграційними потоками Сполучених Штатів Америки.



2

Історія спалахів коронавірусу

- **SARS-CoV-1.**

2003 рік. Гуандун, Китай. 26 країн, близько 10000 інфікованих.

- **MERS-CoV.**

2012 рік. Саудівська Аравія та Країни Близького сходу. 30% смертності.

- **SARS-CoV-2.**

2019 рік. Ухань, Китай. Досі досліджується.

3

SARS-CoV-2

Одноланцюговий РНК-вірус, сімейства коронавірусів, який викликає гострий респіраторний синдром коронавірусів

COVID-19

Тяжке респіраторне захворювання, викликане вірусом SARS-Cov-2

4

Епідеміологічні моделі

Регресійні моделі

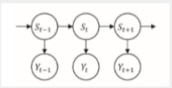
Є одним з найпопулярніших методів прогнозування вірусних захворювань. Основною задачею регресії являється знаходження функціональної залежності показниками захворюваності та факторами, які впливають на захворюваність, задля чого відновлюються та формуються невідомі параметри.

Модель Серфлінга:

$$\hat{y}_t = \gamma e^{\sum_{j=0}^{p} a_j t^j + \sum_{j=0}^{k} (\beta_{2j-1} \sin\theta_j + \beta_{2j} \cos\theta_j) + \rho_1 x_t^1 + \rho_2 x_t^2 + \cdots}$$

Баєсівські мережі

В вірусології Баєсівські мережі активно стають популярними саме у вигляді простої форми прихованих марковських моделей.



Експоненційне згладжування

Для сезонних захворювань використовують так зване потрійне експоненційне згладжування.

Яскравим прикладом потрійного експоненційного зважування є адитивна сезонна модель Хольтера-Вінтерса:

$$l_t = \alpha(y_t - s_{t-p}) + (1-\alpha)(l_{t-1} + r_{t-1}),$$
$$r_t = \gamma(l_t - l_{t-1}) + (1-\gamma) r_{t-1},$$
$$s_t = \delta(y_t - l_t) + (1-\alpha) s_{t-T},$$
$$\hat{y}_{t+h} = l_t + hr_t + s_{t-T+h}.$$

Калманівські фільтри

Фільтри Калмана користуються широкою використовуються в області інженерії та економетрики, і тільки починають набувати популярності в прогнозуванні епідеміологічних процесів. В загальному вигляді епідеміологічні процеси для нашої моделі можна записати в такому вигляді:

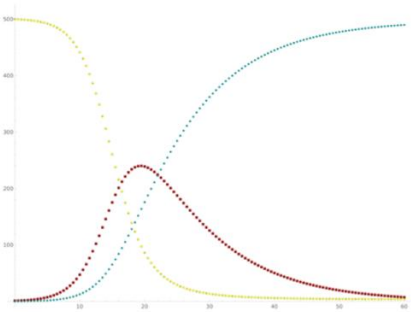
$$x_t = Ax_{t-1} + w_t,$$
$$y_t = Hx_t + Df_t + v_t,$$

5

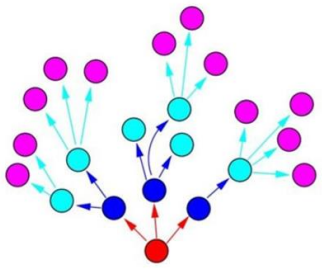
Модель SIR

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta IS}{N}, \\ \dfrac{dI}{dt} = \dfrac{\beta IS}{N} - \gamma I, \\ \dfrac{dR}{dt} = \gamma I, \end{cases}$$

$$S(t) + I(t) + R(t) = const = N$$



6



$$S(t) = S(0)e^{-\frac{R_0\left(R(t)-R(0)\right)}{N}},$$

Число репродукцій R

$$R_0 = \frac{\beta}{\gamma}$$

$$\frac{dI}{dt} = (R_0\frac{S}{N} - 1)\gamma I$$

$$R_0 > \frac{N}{S(0)}$$

7

Модель SIR-F

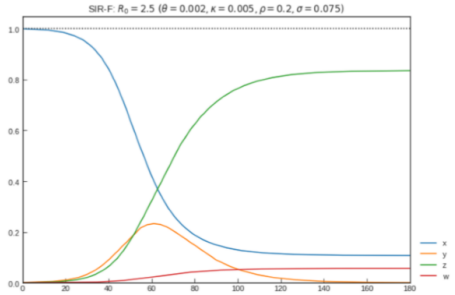
$$\begin{aligned} \mathrm{S} &\xrightarrow{\beta I} \mathrm{S}^* \xrightarrow{\alpha_1} \mathrm{F} \\ \mathrm{S}^* &\xrightarrow{1-\alpha_1} \mathrm{I} \xrightarrow{\gamma} \mathrm{R} \\ \mathrm{I} &\xrightarrow{\alpha_2} \mathrm{F} \end{aligned}$$

$$\frac{dS}{dT} = -\frac{\beta SI}{N},$$

$$\frac{dI}{dT} = \frac{(1-\alpha_1)\beta SI}{N} - (\gamma + \alpha_2)I,$$

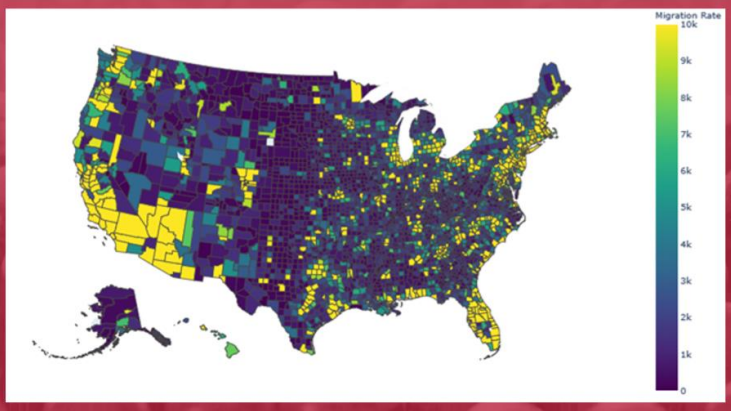
$$\frac{dR}{dT} = \gamma$$

$$\frac{dF}{dT} = \frac{\alpha_1 \beta SI}{N} + \alpha_2 I$$



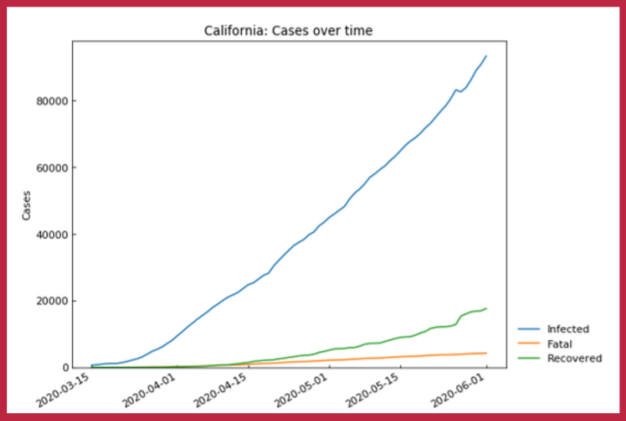
8

Рейтинг міграції



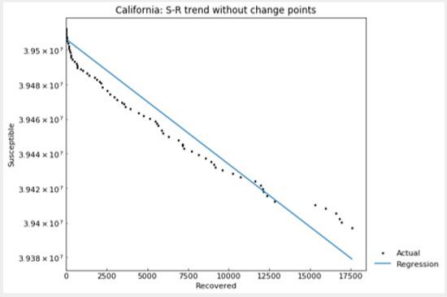
9

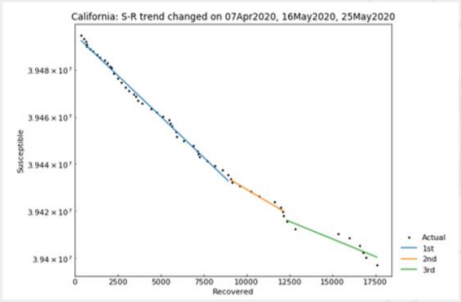
Моделювання епідемії



10

Тренд Аналіз





11

Епідеміологічні показники

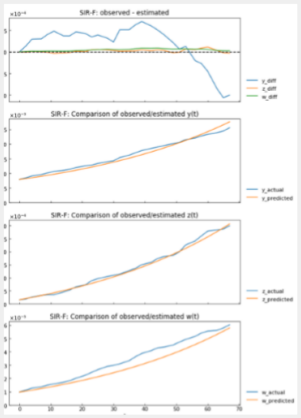
$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log_{10}(A_i+1)-\log_{10}(P_i+1))^2}$$

	Type	Start	End	Population	ODE	tau	theta	kappa	rho	sigma	Rt	alpha1 [-]	1/alpha2 [day]	1/beta [day]	1/gamma [day]	RMSLE
1st	Past	06Apr2020	10May2020	39512223	SIR-F	111	0.025055	0.000052	0.003728	0.000455	7.17	0.025	1470	20	169	0.336508
2nd	Past	11May2020	22May2020	39512223	SIR-F	111	0.050672	0.000046	0.002293	0.000428	4.59	0.051	1684	33	180	0.0768706
3rd	Past	23May2020	01Jun2020	39512223	SIR-F	111	0.038015	0.000042	0.002209	0.000419	4.60	0.038	1831	34	183	0.0872706
4th	Future	02Jun2020	09Jun2020	39512223	SIR-F	111	0.038015	0.000042	0.002209	0.000419	4.60	0.038	1831	34	183	-

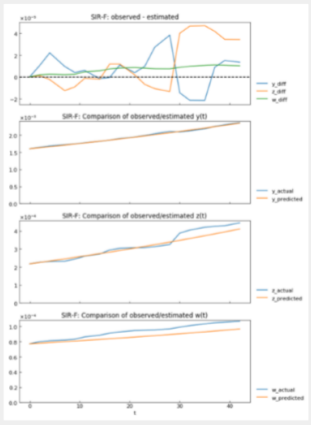
12

Точність моделі

Перша фаза

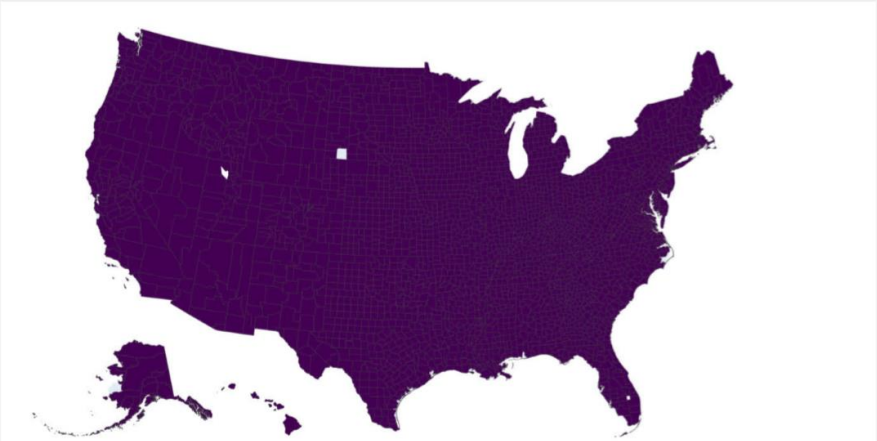


Третя фаза



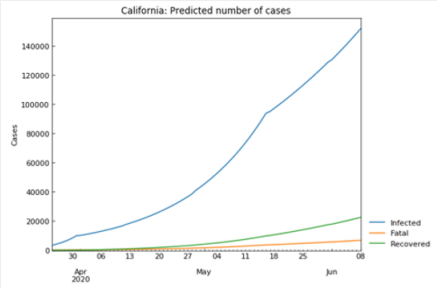
13

Розповсюдження вірусу



14

Результати моделі



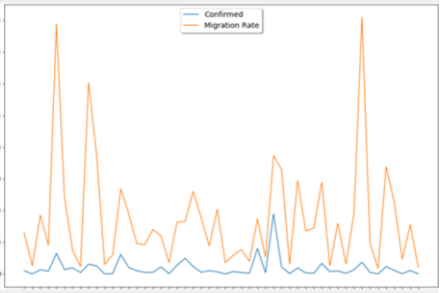
101	02Jun2020	California	121884	4564	117299	0
102	03Jun2020	California	124742	4718	120024	0
103	04Jun2020	California	127687	4875	122812	0
104	05Jun2020	California	127687	4875	122812	0
105	06Jun2020	California	130700	5035	125664	0
106	07Jun2020	California	133783	5200	128583	0
107	08Jun2020	California	136937	5368	131569	0

15

Оцінка результатів

КОРЕЛЯЦІЙНЯ МАТРИЦЯ

	__Estimate	Confirmed	Fatal	Recovered	Infected
__Estimate	1.000000	0.657143	0.564082	0.071336	0.550879
Confirmed	0.657143	1.000000	0.766531	0.131762	0.691459
Fatal	0.564082	0.766531	1.000000	0.083086	0.568860
Recovered	0.071336	0.131762	0.083086	1.000000	-0.038653
Infected	0.550879	0.691459	0.568860	-0.038653	1.000000



	State	1/beta [day]
0	Idaho	66
1	West Virginia	63
2	Florida	52
3	Wyoming	51
4	Washington	47
5	Tennessee	42
6	South Dakota	40
7	Connecticut	37
8	Oregon	37
9	New York	34

	State	Estimate
39	Texas	1621404
4	California	1580553
8	Florida	1209244
9	Georgia	750873
28	New York	748434
42	Virginia	678138
29	North Carolina	659778
31	Ohio	591360
34	Pennsylvania	579098
12	Illinois	538151

16

Висновки

На основі проведеного аналізу існуючих моделей та за обраною темою дослідження був створений програмний додаток для аналіз показників захворювання COVID-19 та аналізу впливу міграційних потоків на процес та швидкість розповсюдження захворювання на прикладі Сполучених Штатів Америки.

В ході даної роботи було створено програмний продукт для аналізу та симуляції епідеміологічних процесів захворювання COVID-19 та знаходження зв'язку між ними та міграційними потоками на внутрішньому рівні Сполучених Штатів Америки.

В процесі дослідження епідеміологічної сфери були розглянути основні моделі для передбачення, аналізу та симуляції різнотипних захворювань та вірусів, починаючи від самих перших спроб створення моделей, до найбільш популярних та найбільш використовуваних алгоритмів дослідження.

17

Дякую за увагу!